AN INVESTIGATION OF THE RASCH MODEL IN ITS APPLICATION TO FOREIGN LANGUAGE PROFICIENCY TESTING

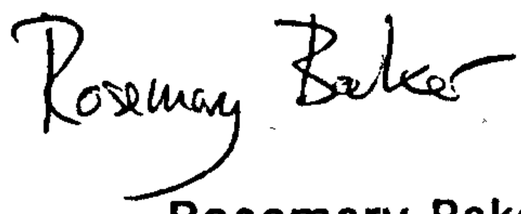
Rosemary Lilian Baker

Ph.D.
University of Edinburgh
1987

For Susanne and John Cooper

DECLARATION

I declare that this thesis consists entirely of my own work, and that it has been composed by myself.



Rosemary Baker

ABSTRACT

The main theme of this study is the use of the Rasch model for dichotomously-scored items in the analysis of second/foreign language proficiency test data. Analytic procedures deriving from this model are applied to response data from English proficiency tests of two different types: (i) a cloze-type test, which embodies the notion of proficiency as being measurable by a single, global test, and (ii) the objectively-scored sections of the English Language Testing Service (ELTS) test, in which proficiency is viewed as being divisible into sub-components, each measured by a separate subtest. The total numbers of testees involved are 854 and 1,503 respectively.

The theoretical background relating to item response models is first explained, via a discussion of traditional procedures for the analysis of test data. The relationship between the Rasch model and other response models of similar mathematical form is considered, and further details of its operation provided.

The results of the Rasch analyses are compared with those from traditional analyses of the same data. The Rasch statistics are shown to be more informative, and therefore preferable, on several counts.

Further investigations are carried out on both data sets, in order to assess the fit between model and data, to check for possible violations of specific model assumptions, and to check for expected model features. For both the cloze-type data and the ELTS data (analysed in separate subtests), observed and expected item characteristic curves show reasonable conformity, though with some instances of serious misfit in both cases. No evidence for departure from unidimensionality is found for the cloze data, but there is some indication that ELTS modules, when combined with the General components, may vary in their departure from unidimensionality.

ACKNOWLEDGEMENTS

I am indebted to my supervisor, Dr Clive Criper, for making available to me the language test data used in this study and for the advice which he provided during the course of this work.

Particular thanks are due to Alastair Pollitt, for introducing me to Rasch theory, showing me how to carry out (and interpret the results of) Rasch analysis, and contributing greatly to my understanding of the area. I also gratefully acknowledge use of Mr Pollitt's computer program for traditional item analysis.

Scholarships from the Institute for Applied Language Studies and from the Faculty of Arts, University of Edinburgh made it possible for me to pursue this research.

The appearance of this thesis owes much to the word-processing expertise of Shirley Laird and Gus Macdonald, and I am very grateful to both for the care, efficiency and speed with which they worked on the final version. I should like also to thank Gibson Ferguson and Ian McGrath for their encouragement and practical assistance.

Finally, I should like to express my gratitude to my husband, Graham Baker, for his unfailing encouragement, support and patience at all stages in the preparation of this thesis.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1
INTRODUCTION

It might appear, from the remarks of Hambleton and Murray (1983:71), that item response theory (IRT) is now well known and generally accepted by those concerned with the development of educational tests:

> "The many applications appear to be so successful that discussions of IRT have shifted from consideration of model advantages and disadvantages compared to classical test models to consideration of topics such as model selection, item and ability parameter estimation, and methods for determining goodness of fit."

As far as the area of second/foreign language testing is concerned, however, it is only within the last four years or so that mention of item response theory has been made in the literature: further evidence of its comparatively recent arrival in language testing circles may be seen in the fact that introductory seminars for language testers were organised in 1985, both by the Educational Testing Service in Princeton, and by the British Council in London.

Applications of IRT-based procedures, and in particular of those deriving from the Rasch model, are now being reported with increasing frequency in the language testing literature. This tends to give the impression that these procedures, and the body of psychometric theory upon which they are based, are now familiar to language testers in general: however, the constraints imposed by the length of articles are such that only the briefest of introductions to the theoretical background can usually be given, and it is by no means certain that such familiarity may be assumed.

One of the aims of this study, therefore, is to present a more explanatory account of the theoretical background of IRT than is possible in most published work intended for language testers. This is introduced via a consideration of the traditional procedures for test analysis, which may safely be assumed to be already thoroughly familiar; it is hoped that in presenting the discussion in this way, the relationships between the traditional and IRT-based approaches will be demonstrated more clearly than is sometimes the case.

This account is for the most part concerned with the three most widely-mentioned item response models for dichotomously-scored items; in the latter part of Chapter 2, however, attention narrows to one of these, the Rasch model, since this is the model upon which the analyses of test data presented in later chapters are based.

In Chapter 3, reported applications of Rasch analysis in the area of second/foreign language testing are surveyed, and, on the basis of these, a number of issues for investigation identified.

The main set of data analyses, in which traditional and Rasch procedures are applied to data from a measure of 'overall' language proficiency (in the form of a cloze-type test), is discussed in Chapter 4. The information obtained using the two different approaches is summarised, interpreted and compared, with the aim of further clarifying the relationship between the two. Further investigations of the cloze-type data are then described: these are based on methods suggested in the IRT literature as ways of assessing the extent to which the data conform to the chosen response model, and for checking for possible violations of specific model assumptions. The issues considered here include the dimensionality of the data set, the sample-independence of the Rasch difficulty estimates, and the test-independence of the Rasch ability estimates.

In Chapter 5, similar methods of analysis are applied to data from three subtests of a test battery (the ELTS test). Since this test represents a view of language proficiency as being composed of many different sub-components, it provides a potentially interesting point of comparison for the results obtained using the cloze-type test.

CHAPTER 2
TRADITIONAL AND RASCH APPROACHES TO TEST ANALYSIS

2.1 INTRODUCTION

The traditional and Rasch approaches to the analysis of test data both have their roots in theories of testing developed within the context of psychological and psychometric research. The purpose of this chapter is to set the two approaches against this background, to outline the development and formulation of their theoretical bases, to consider the assumptions made in each, and to discuss the implications of each for practical testing work. Since the main focus of this study is the use of Rasch analysis, this approach is presented in greater detail; discussion of the traditional approach is included largely for purposes of comparison.

2.1.1 Background to Psychological Measurement

Research in psychology has for over a century been concerned both with the search for generally applicable principles of human behaviour, and with the study of individual differences (Tyler & Walsh, 1979:26), and the main aim in psychometric research, as defined by Bock (1983:113), has been "... to describe persisting characteristics of individual subjects as dependably as possible."

Attempts have thus been made to identify, isolate and measure a wide variety of (hypothetical) constructs, including e.g. anxiety, intelligence, attitudes and motivation. Such constructs are not, of course, directly observable or measurable: their levels can only be inferred, through the reactions of individuals to given stimuli (Samejima, 1983:159), which in practice may mean their responses to test items. A test, then, can be seen as "... a procedure designed to elicit certain behaviour from which one can make inferences about certain characteristics of an individual" (Carroll, 1968:46), and measurement in this context can be defined as "... the process of assigning numerical values to a person's performance in accordance with specified rules" (Brown, 1976:2).

Although, as Thorndike and Hagen (1977:9) observe, measurement in any field involves the three common steps of (i) defining the attribute to be measured, (ii) determining a set of operations for making the attribute manifest, and (iii) establishing procedures for translating observations into quantitative statements, the measurement of psychological characteristics is frequently contrasted with physical measurement, and shown to be more problematic.

Thorndike and Hagen (1977:10) mention, for example, the difficulty of defining a psychological attribute such as 'intelligence' as compared with the definition of a physical attribute such as 'length', and note the lack of consensus as to the appropriate procedures for measuring it. Lord and Novick (1968:13–14) set out differences in the conditions under which the two types of measurement are typically carried out: they note (a) that in the physical sciences the same measurement can normally be repeated several times, whereas in psychological measurement such repetition might cause the person's responses to change because of fatigue or practice effects, and (b) that in the physical sciences inferences are usually made about one object or event at a time, while in psychological measurement the concern is often to make inferences both about the individuals in a group and about the group as a whole. Brown (1976:9) points out that in psychological measurement more variables have to be controlled than in physical measurement, and many writers (e.g. Ebel, 1972:39; Stanley, 1972:60–61; Thorndike & Hagen, 1977:11–13; Guilford & Fruchter, 1978:23–24) observe that in physical measurement there are well-defined scales with equal units, while in psychological measurement there are no generally accepted units and scales, and equality of units is considerably more difficult (or indeed impossible) to establish.

A major task in psychometrics has therefore been to try to overcome these difficulties, and to place the measurement of psychological attributes on a sound theoretical basis. The measurement models and associated statistical procedures described in this chapter have been developed with this aim.

2.1.2 Use of Psychometric Methods in Educational Testing

Procedures developed originally to aid in the construction of psychological measures have come to be applied also in educational testing; indeed, the use of psychometric methods in the development of educational tests is advocated in numerous textbooks (see e.g. Ebel, 1972; Payne & McMorris, 1975; Brown, 1976; Thorndike & Hagen, 1977; Nitko, 1983). Such methods are described in most handbooks concerned specifically with the construction of second/foreign language tests (e.g. Harris, 1969; Heaton, 1975; Allen & Davies, 1977; Oller, 1979; Henning, 1987).

A question sometimes raised, however, is whether this transfer of methods is appropriate, given that in educational testing it is frequently even more difficult than in psychological testing to isolate and measure a single attribute: in an educational setting there will usually be many additional, irrelevant variables

which are impossible to control. Furthermore, the 'measures' typically required in an educational context are of multidimensional abilities and achievements rather than of one dimension at a time. Brown (1980:20) suggests that the criteria for psychological measurement may not be consistent with those relating to the assessment of attainment, and Goldstein (1980a:211) argues that educational measurement and psychological measurement are quite different activities: he expresses concern that traditional psychometric theory, which takes unidimensionality for granted, "... may come to be used too automatically in an inappropriate context."

Choppin (1981:213-5), on the other hand, considers that measurement has a valuable role to play in education, particularly in the diagnosis of individual pupil difficulties, the monitoring of standards of achievement, and curriculum evaluation. He emphasises (p.205-7) that a clear distinction must be drawn between measurement, which implies a quantification of something, and is therefore necessarily unidimensional, and operations such as examination or assessment, which may include several independent dimensions (and in so doing provide less easily interpretable results).

The application of procedures deriving from psychometric theory in the development of language tests has also met with criticism in recent years. Morrow (1979), for example, regards the quantification implied by the use of these methods as inappropriate in the assessment of language proficiency, and Cziko (1983) views psychometric methods as being based too heavily on a norm-referenced interpretation of scores and on the maximising of individual differences, approaches which are considered undesirable in some quarters.

Despite the misgivings expressed, the need remains in language testing, as in any other area of educational testing, for analytic tools which can be used to aid in the development of test instruments. This study is concerned with two such sets of analytic procedures: traditional test analysis, which has its roots in classical test theory (CTT), and Rasch analysis, which derives from a measurement model seen as belonging to a family of models often referred to collectively as 'Item Response Theory' (IRT).

As Hulin, Drasgow and Parsons (1983:67) observe, CTT and IRT might best be viewed as partially overlapping rather than as rival theoretical frameworks. There are, however, important differences between them, and it is the purpose of Sections 2.2 and 2.3, by considering the two frameworks in turn, to make these clear.

2.2 CLASSICAL TEST THEORY

The foundation for much of classical test theory was provided by Charles Spearman's conception of an observed test score as being composed of a true score plus an error component (Thorndike, 1982a:3), and as Gulliksen (1950:1) explains, most of the basic formulae which have proved to be particularly useful in this theory appeared in Spearman's work in the early 1900s. Subsequent developments in classical test theory are set out in textbooks written by Gulliksen (1950) and by Lord and Novick (1968), both of which are seen as representing major contributions to the field (Hambleton & van der Linden, 1982:373). Spearman's initial formulation remains central, however.

2.2.1 The Basic Model

In the basic model of classical test theory, then, an individual's observed score, X, on a given test is expressed as:

$$X = T + E$$

where T = the true score and E = the error of measurement. Guilford and Fruchter (1978:409) note that the true score is conceived of as being the score that would be obtained if a perfect measuring instrument were applied under ideal conditions; an operational definition of true score, however, would be the mean score obtained by the person over a very large number of repeated administrations of the test.

The true score is viewed as remaining constant over all administrations, and over all parallel forms of the test (Thorndike, 1982a:4). The error component is included in the model to account for any non-systematic variation observed in a person's scores from one occasion to another, or from one form of the test to another. Errors of measurement are seen as fluctuating randomly around the true score, and the mean error of measurement is assumed to be zero (Stanley, 1972:64; Guilford & Fruchter, 1978:409). Since variation in observed scores is attributed to random error, errors of measurement are assumed to be completely independent of true scores.

2.2.2 Reliability and Error of Measurement

The traditional approach to test reliability is based on this relationship between true scores, observed scores and errors of measurement. As Linn and Werts (1979:54) explain, it follows from the basic assumptions of classical test theory that the variance of the scores obtained for a group of individuals on a

given measure is equal to the variance of their true scores plus the variance of random error:

$$\sigma_X^2 = \sigma_T^2 + \sigma_E^2 .$$

The reliability of the set of measurements is intended to indicate the proportion of the variability in observed scores attributable to variability in true scores, and is defined as the ratio of true score variance to observed score variance:

$$\text{reliability} = \frac{\sigma_T^2}{\sigma_X^2}$$

(see e.g. Guilford & Fruchter, 1978:410; Linn & Werts, 1979:54; Krźanowski & Woods, 1984:3). A perfectly reliable measure would be one on which the differences in observed scores reflected only differences in true scores, and where error played no part; if this were the case, the above reliability coefficient would take the value 1.

Since the relative contributions of true score and error variance to observed score variance are unknown, further steps are necessary before reliability can be estimated. The approach usually taken is to define a second measure on which all individuals are assumed to have the same true scores as on the original measure, but for which the errors of measurement are independent (Linn & Werts, 1979:54-55). A further assumption made is that the variances of the errors of measurement will be equal for the two measures. It can then be shown that the reliability of the measures is equal to the correlation between them (Linn & Werts, 1979:55; Krzanowski & Woods, 1984:6).

In practice, this second set of scores can be obtained in a number of ways. One suggested procedure is to administer the same test twice, and to calculate a 'test-retest' reliability coefficient. For this coefficient to be interpretable, steps must be taken to ensure that the second set of scores is not influenced by practice or memory effects, or, in the case of a longer interval between administrations, that no learning, and hence no change in true scores, has taken place in the intervening period (Ebel, 1972:412; Krzanowski & Woods, 1984:6). Clearly, it is impossible to establish with any certainty the required stability of true scores and randomness of measurement errors. A further problem, pointed out by Ebel (1972:412), is that this procedure gives no indication of the differences in scores that might be obtained if a different sample of items from the (usually very large) population of possible items had been selected.

The 'parallel', 'alternate' or 'equivalent' forms method, in which the sets of

scores are obtained from two different but equivalent versions of the test, has been used as an alternative. One practical problem which arises here is that many educational achievement tests, particularly those for classroom use, are produced singly (Ebel, 1972:412). The greatest objection raised, though, concerns the requirement that the different forms of the test be equivalent. Krzanowski and Woods (1984:6) caution against believing, a priori, that an individual's true score will be the same for the different forms, noting that a new version might well introduce additional variability into the results. As Linn and Werts (1979:55) point out, "... the test publisher's reference to the correlation between two alternate forms of a test as a reliability coefficient rests on a series of rather strong assumptions." They recommend, as a minimal check for equivalence, a comparison of the means and variances of the observed scores on the two forms; if these are not the same for the two measures, they cannot be considered parallel.

The 'split-halves' method of estimating reliability overcomes the problem of needing two different versions of a test by treating the two halves of a single test as though they were separate but equivalent. It also avoids the problem of possible changes in true score, by requiring only one administration; the reliability estimate obtained for the half test can be 'stepped up' to indicate the reliability of the full-length test by using a special version of the Spearman-Brown formula (Guilford & Fruchter, 1978:426). However, Ebel (1972:414) observes that a split-halves reliability estimate may be influenced by the way in which the test is divided, since some divisions are likely to yield more closely equivalent halves than others.

The test-retest and parallel forms methods can both be viewed as ways of assessing the stability of measurement, over separate administrations of a test and across different forms of a test respectively. The split-halves method, on the other hand, in being concerned with the degree of correlation between parts of the same test, is more closely related to the set of methods usually said to estimate the internal consistency (or homogeneity) of a measure.

Several such methods have been developed, largely, according to Guilford and Fruchter (1978:427), as a result of dissatisfaction with the arbitrary splitting of tests into halves. Those most commonly discussed are Cronbach's coefficient alpha, which is a general formula, and the Kuder-Richardson formula, which requires less information, but which is appropriate only if all the items in the test are of approximately equal difficulty (Guilford & Fruchter, 1978:428).

These methods, which require only a single administration of a single test, are consistent with the classical definition of reliability as the ratio of true variance to total variance. As Guilford and Fruchter (1978:430) explain,

> "The total variance of a test can be conceived as a sum of the variances and covariances of its parts. The true variance of a test is contributed by its covariances, to which both the item variance and item intercorrelations are important contributors."

They also note that the item variances contribute to internal consistency only by virtue of entering into the covariance terms. Thus internal consistency reliability is determined essentially by the intercorrelation of the items: the larger these intercorrelations, the greater the internal consistency (Guilford & Fruchter, 1978:424).

According to Thorndike and Hagen (1977:82), the internal consistency reliability coefficient indicates "... the degree to which all of the items measure a common characteristic of the person." However, Nunnally (1978:229-230) points out that the size of the coefficient depends not only on the average correlation among the items, but also on the number of items: the coefficient can be increased by incorporating additional suitable items into the test.

The size of the coefficient is also affected by the degree of variation in item difficulty. As Guilford and Fruchter (1978:424) observe, a wide difference in the proportions of testees answering two given items correctly will restrict the correlation between the items. Thus test homogeneity will not be all that is reflected in the reliability coefficient, as is clear from the following:

> "Two items very far apart in difficulty might correlate less than .20 even when each measures the same thing and measures it well." (Guilford & Fruchter, 1978:424)

Ebel (1972:433) also mentions the effect of variation in item difficulty on the reliability coefficient, noting that the wider the variation, the more serious the likely underestimation of reliability.

A related factor affecting the size of the coefficient is the dispersion of scores in the group tested. The coefficient obtained from a group which is relatively homogeneous with respect to the measured attribute will be lower than for a more heterogeneous group (Guilford & Fruchter, 1978:431).

Thus internal consistency reliability will appear highest when items are of equal difficulty, thereby allowing maximum item intercorrelations, and of median difficulty, resulting in maximum item variances (Guilford & Fruchter, 1978:430).

Indeed, Guilford and Fruchter state (p.427) that the assumptions of the Kuder-Richardson methods call for items of approximately equal difficulty and equal intercorrelaton. (A modification suggested by Horst (1953) allows for variation in item difficulty, however.) They further note that these methods are most appropriately applied to homogeneous tests, and that their use is "entirely precluded" for tests of speed rather than power (p.430).

A number of general points can be made regarding the use of the traditional procedures for estimating reliability outlined here. Concern is expressed by Krzanowski and Woods (1984:13) that language testers often seem content to make use of the formulae given in the literature, and to report the results, without taking account of the likely error of the estimates. Such estimates are also often reported as though they indicated the reliability of the test itself, irrespective of the particular circumstances in which it was administered. Guilford and Fruchter (1978:408) are careful to point out that reliability is estimated for a given set of measurements, not for the measuring instrument itself. They write:

> "It can rarely be said of any instrument, whether a test or some other device, that *the* reliability of the device is of a certain value .... One should speak of the reliability of a certain instrument applied to a certain population under certain conditions."

They therefore emphasise (p.430) that reliability coefficients "... must be interpreted in a relativistic manner".

A further important point, arising from an inherent feature of classical reliability theory rather than from its misuse, is that it implies that stability or consistency of measurement is the same for all persons in the group tested. This is also reflected in the standard error of measurement (the estimated standard deviation of the errors of measurement, expressed in test score units), which, like the reliability coefficient, is reported as a single, global statistic for the whole group.

2.2.3 Traditional Item Statistics

Hulin, Drasgow and Parsons (1983:68) observe that strictly speaking, item parameters are not defined in classical test theory, but that those who develop measuring instruments in accordance with this theory typically make use of an item difficulty (or 'facility') statistic, which is usually simply the proportion of correct responses to each item, and an index of item discriminating power, which indicates the degree to which each item distinguishes between high- and

low-scoring persons. As Thorndike (1982a:6) explains, the pretesting and statistical analysis of test items became accepted practice in the U.S.A. during and after the First World War, when the use of multiple-choice and true-false items first became popular. He notes that many different statistical indices of discrimination were suggested during the 1920s and 30s, but that by 1940 the biserial or point biserial correlation between item and total test score had been generally adopted. The importance of these item statistics was not only that they could help the test developer to identify items which had been poorly constructed, but that once incorporated into psychometric theory, they were of use in the selection of items to produce tests with certain known properties: guidelines as to the levels of difficulty and discrimination that should be aimed for are provided in most textbooks on testing (e.g. Ebel, 1972; Brown, 1976; Nitko, 1983).

Clearly, though, the 'difficulty' of an item as reflected by the proportion of correct responses given by a group of testees will vary according to whether the testees are of high or low levels in the attribute being measured. As Lord (1980:35) explains,

> "Proportion of correct answers in a group of examinees is not really a measure of item difficulty. This proportion describes not only the test item but the group tested."

Thus the traditional item difficulty index is sample-dependent: the values obtained will remain stable only for groups of similar levels. Even if the facility values are converted to a standard-score scale such as that described by Guilford and Fruchter (1978:458-9), there will still be systematic differences depending on the levels and ranges of the groups on whose responses the calculations are based.

The commonly used discrimination indices are coefficients of correlation between a dichotomous variable (correct vs incorrect response to the item) and a continuous variable (the total number of items which the person answers correctly). The point biserial correlation, which is the Pearson product-moment correlation between these two variables (Lord & Novick, 1968:336), is more generally applicable than the biserial correlation, as it does not involve restrictive assumptions (Guilford & Fruchter, 1978:310). However, as is noted both by Gulliksen (1950:393) and by Lord and Novick (1968:341-2), this index has been found to vary systematically with item difficulty, and hence with the level of the group tested. It also varies according to the distribution of the measured attribute within the group, tending to be higher if the group is heterogeneous

rather than homogeneous with respect to the attribute. This results from the "... well-known effect of homogeneity on correlation coefficients" referred to by Lord and Novick (1968:354).

The biserial correlation is described (e.g. by Lord & Novick, 1968:337; Hulin et al., 1983:237-8) as the correlation between a hypothesised continuous variable underlying the correct-incorrect dichotomy imposed in scoring the items, and the continuous variable represented by the total test score. According to Nunnally (1978:136), it provides an estimate of the product-moment correlation that would be obtained if the dichotomised variable were normally distributed; it is said by Hulin et al. (1983:76) to correct the product-moment correlation between item score and total test score.

Although it was hoped that the biserial would demonstrate stability from group to group (Lord & Novick, 1968:341), this has not always proved to be the case: Gulliksen (1950:393), for example, notes that although the biserial is not prone to systematic bias in theory, bias has nevertheless been found in practice. Wood (1976:255), too, reports that biserials for ostensibly equivalent groups have shown considerable variation, beyond that which might be predicted from sampling theory.

As was mentioned above, the biserial correlation assumes that the hypothesised underlying variable is normally distributed; Guilford and Fruchter (1978:307) warn that marked departures from normality may lead to erroneous results. They further note that while it is not necessary for the continuous variable to be normally distributed, it should be unimodal and roughly symmetrical. Other assumptions, set out by Hulin et al. (1983:238), are (a) that the observed dichotomous variable results from imposing on the hypothesised underlying variable a threshold which separates those who will answer correctly from those who will answer incorrectly, and (b) that the regression of the observed continuous variable onto the hypothesised underlying variable is linear. They acknowledge (p.76) that the assumptions made are often violated, but consider the effects of this to be less serious than the effects of variation in item difficulty on the point biserial correlation. Nunnally (1978:136-7), on the other hand, issues a strong warning against the use of the biserial correlation coefficient, both on the grounds that it is always higher than the point biserial coefficient, in some cases misleadingly so, and because of the considerable error which can result when the assumption of normality is not met.

A simple method for obtaining a discrimination index (often referred to as an

'E' index) is described in several of the language testing textbooks, including Harris (1969:106), Heaton (1975:174) and Allen and Davies (1977:187-9). This involves calculating for each item the proportion of correct responses given by the highest-scoring half, third or quarter of the sample, and subtracting from this the proportion of correct responses given by a similar subsample drawn from the lowest-scoring persons. Again, the results will be affected by the particular distribution of the measured attribute within the sample tested, and so may vary from one sample to another.

This feature of sample-dependence has long been seen as a disadvantage of the traditional procedures for item analysis. Gulliksen (1950:367-70) describes a number of attempts which have been made to overcome this problem, but indicates that these have not succeeded when he then writes, in a much-quoted passage (see e.g. Choppin, 1976:237; Gustafsson, 1977:1):

> "A significant contribution to item analysis theory would be the discovery of item parameters that remained relatively stable as the item analysis group changed; or the discovery of a law relating the changes in item parameters to changes in the group." (Gulliksen, 1950: 392)

Lord and Novick (1968:328) also emphasise the importance of item parameters which remain invariant from one group of examinees to another, noting that in practical testing work there are often systematic differences between pretesting groups and the groups with whom a test is later used. Unless the pretesting group is representative of these groups, the item statistics obtained are likely to be limited in their usefulness, and may be misleading.

It should be noted that for the item statistics discussed here, and for the estimates of internal consistency mentioned in the previous section, the assumption is made that people who answer items incorrectly, or who fail to answer them at all, do so because of an insufficient level in the attribute being measured, and not as a result of not having had time to attempt all the items. If this assumption does not hold, and some testees are prevented by time limits from reaching items that they would otherwise have been able to answer correctly, some of the item scores will be artificially low, and the other statistics deriving from these will be affected.

It is similarly assumed that where people answer items correctly, this reflects a certain level of knowledge, skill or ability in whatever is being tested, and is not merely the result of correct guessing. The effect of correct guessing, particularly if it occurred on a large scale, would be the artificial inflation of some of the

item scores, with its consequent effects on the other statistics. (A method for 'correcting' multiple choice item scores for chance success is described by Guilford and Fruchter (1978:460-1); however, this technique is based on the notion that those who do not know the correct answer make random guesses among the available choices, and this is thought unlikely to be the case.) Although these assumptions are not always stated explicitly, they are implicit in the use of the number-correct item scores which form the basis of item analysis procedures.

2.2.4 Traditional Person Scores and Scales

Following traditional testing procedures, person scores on a test composed of dichotomously-scored items are usually reported simply as the number (or percentage) answered correctly, or in the form of a score derived from this in one of several possible ways. Use of number-correct person scores, and indeed of the various derived scores, rests on a number of assumptions which, again, are not always made explicit, and which are therefore sometimes not fully recognised.

Pollitt (1979:59) points out that to count up and report the total number of correct responses is to assume that the test is unidimensional, i.e. that the items are all in some sense measuring the same thing. The point raised by Choppin (1981:207), concerning the difficulty of interpreting the total score on a multidimensional test, has already been mentioned (see Section 2.1.2), and although such scores are often treated as though they represented coherent measurement of something, Preece (1980:209) would argue that they are "largely meaningless".

Furthermore, use of number-correct person scores implicitly assumes that the items discriminate equally (Pollitt, 1979:60; Wright, 1977b:220), and, unless some form of differential weighting is employed, that each item represents an equivalent unit of measurement, i.e. that placed together, the items form the basis for an equal interval scale. However, it is widely acknowledged (see e.g. Lord & Novick, 1968:22; Ebel, 1972:83; Brown, 1976:11) that psychological and educational tests cannot generally be said to measure on anything more than ordinal scales, since, as Thorndike and Hagen (1977:15) explain, "... the equality of units cannot be established in any fundamental sense." Thus although numerically equal differences in scores are commonly treated as evidence of equal differences in people's standing on the measured attribute, and various mathematical operations are routinely performed on test scores, such practices

may not be appropriate: the only procedures which are, strictly speaking, permissible when measurement is on ordinal scales are statistical procedures based on ranks (Tyler & Walsh,1979:6) and transformations which preserve rank order (Lord & Novick, 1968:21).

It is perhaps not always appreciated that to compute, for example, the mean and standard deviation for a set of scores is to assume that the test measures on an interval scale. For practical purposes, it has been found necessary to make this assumption, however, and this is usually justified with reference to the advantages to be gained from the point of view of the statistical methods which can then be used on the data (see Guilford & Fruchter, 1978:24; Tyler & Walsh, 1979:9), or in terms of the usefulness or plausibility of the outcome. Lord and Novick (1968:22) consider that the assumption is justified if the resultant scale proves to be a good empirical predictor of a relevant criterion, and both Brown (1976:11) and Tyler and Walsh (1979:9) take the view that confidence in the correctness of the assumption can be judged by the extent to which the results and conclusions based on it appear reasonable.

Nunnally (1978:24-26) objects strongly to the suggestion that most psychological tests yield only rank orderings of people, and represent no higher form of measurement. He argues that the absence of unequivocal evidence to support the assumption of an interval scale does not necessarily mean that only an ordinal scale is present, and claims that it is in any case not clear what would constitute such evidence. He points out that even attributes such as temperature, time and steam pressure can only be measured indirectly, through measuring their correlates (e.g. the height of a column of mercury, the swing of a pendulum, the movement of a pointer on a gauge), and therefore considers it unreasonable to insist that equality of intervals in the measurement of intelligence, for example, should be established in some more direct way.

Ebel (1972), while accepting that educational test scales have a number of limitations from a technical measurement point of view, does not regard this as a major practical concern. He writes:

> "For inadequate as the scales are, the errors they introduce into educational measurements are far less serious than the errors associated with the definition of the trait to be measured and with the selection and presentation of tasks to be included in the test. The basic problems of educational measurement are not problems of scaling, but problems of test planning and item writing." (Ebel, 1972:83)

Nevertheless, for purposes of reporting, interpreting and comparing scores, the

need has been recognised for scales other than those provided by simple counts or percentages of correct responses, and the traditional solution has been the use of various types of derived scale.

Such scales have been obtained from raw scores both by linear transformations, which retain the characteristics of the original raw score distribution, and by non-linear transformations, which usually result in changes to the original distribution (Gulliksen, 1950:274).

Standard scores, or 'z-scores', which represent a basic linear transformation of the raw score scale, are obtained by expressing raw scores in terms of standard deviation units above or below the group mean (Gulliksen, 1950:268; Thorndike & Hagen, 1977:129). The resultant standard scale has a mean of 0 and a standard deviation of 1. In order to avoid the inconvenience of using decimals and negative numbers, z-scores have often been used as the basis for scales with more convenient means and standard deviations, referred to by Gulliksen (1950:272) as 'linear derived scores'. These are obtained by the transformation:

$$\text{score} = (\text{chosen SD} \times z) + \text{chosen mean}$$

(Gulliksen, 1950:273). It may suit the test constructor's purposes to have, for example, a scale with a mean of 50 and a standard deviation of 10, or, if a broader classification of test performance is required, a mean of 5 and a standard deviation of 2.

Linearly derived scores of the type mentioned above are sometimes said to be on scales of equal units (see e.g. Brown, 1976:11). Thorndike and Hagen (1977:131) point out, however, that since these scaling procedures change the size of the score scale units uniformly throughout the score scale, they do not make the units equal if they were not equal at the outset. As Guilford and Fruchter (1978:478) explain, for the derived scale to be an interval scale, the obtained sample distribution must be the same as the population distribution would be on a scale of equal units: equality of units will not be 'improved' by the scaling procedure.

One of the main uses suggested for standard scores of this kind has been to provide a common scale for the comparison of scores obtained by the same people on different measures (see e.g. Thorndike & Hagen, 1977:130). Guilford and Fruchter (1978:476) caution, however, that accurate comparisons are possible only if two conditions are satisfied, these being (a) that the population in question must have equal means and dispersions in all the attributes measured

by the different tests, and (b) that the shape of the distribution must be similar for the various attributes. Although they consider it "almost certain" that derived scores are more nearly comparable than raw scores, they acknowledge that in performing these common scaling operations, uniformity of means, standard deviations and forms of distribution can often only be assumed.

The most commonly used non-linear transformations of raw scores have been those which result in percentile scores and normalized scores (Gulliksen, 1950:267). Percentile scores simply indicate the percentage of testees in the sample who scored less than any given raw score, but despite the ease with which such scores can be understood, they are felt to have serious disadvantages. Gulliksen (1950:278) notes that they "... cannot legitimately be subjected to the usual arithmetical operations", and observes (p.280) that if they are expressed as averages, or used in the calculation of correlation coefficients, the results will be misleading. More serious, however, is the fact that percentile scores are not comparable from group to group or from test to test (Gulliksen, 1950:280).

Normalized scores are standard scores developed from the percentile ranks corresponding to the raw scores rather than from the raw scores themselves; they further differ from linear derived scores in that the obtained frequency distribution is "... distorted from its original shape into a normal distribution" (Gulliksen, 1950:280). The main justifications for this are, according to Gulliksen (1950:280), that the normal curve has many convenient properties and that many distributions have in any case been found to be normal.

As with the linear derived scales outlined above, it is possible to choose a convenient mean and standard deviation for the normalized scale. Where the mean is set at 50 and the standard deviation at 10, the resultant scale is generally referred to as a 'T scale' (Thorndike & Hagen, 1977:132; Guilford & Fruchter, 1978:478). Another well-known normalized scale is the stanine scale, which uses a mean of 5 and a standard deviation of (almost) 2, and thus has a range of 9 units. ('T scale' is occasionally used to refer to a linear derived scale with a mean of 50 and a standard deviation of 10; most authors reserve this term for the normalized version, however.)

Gulliksen (1950:280) considers the use of normalized scores to be appropriate when there is reason to believe that the attribute in question is normally distributed, and that a non-normal distribution of observed scores was brought about by defects in the test. Guilford and Fruchter (1978:483-4) advocate their

use in cases where it is not known that an attribute is normally distributed, but "... where there is no inhibiting information to the contrary".

Guilford and Fruchter (1978:484) describe normalized standard scores as "more common and meaningful" than those on the original scale. Gulliksen (1950:282) notes, however, that for normalized scores from different tests to be comparable, the groups must be similar in size, and the distribution of extreme scores must be similar for all the tests in the comparison. This last condition is necessary because of the large differences in reported scores that can result from slight differences in grouping at the extremes when using normalized scores (Gulliksen, 1950:281).

All of the raw score transformations mentioned here yield scores which indicate the individual's standing in relation to some group. Indeed, this type of interpretation of test scores has become widely established, and has been reinforced by the setting up of norms for various populations on different measures, an approach which has been particularly prevalent in the U.S.A.. This has been achieved by administering tests to large groups of people selected as representative of the target population, and using the results obtained as a framework for interpreting the scores of those tested subsequently. Whether this process makes use of percentiles or of one of the various types of standard score scale, the scores depend for their meaning on the standardizing samples having been representative of those with whom the test is used. When standardizing samples are not used, or are not available, and the individual's performance is simply assessed with reference to the group with whom he/she was tested, the result will depend on characteristics of that group, most notably on the level and distribution within it of the measured attribute or ability.

Even when scoring does not rest on group-related procedures, as, for example, when number-correct or percentage scores are reported, the scores constitute an index of performance relative to a specific set of items. They are therefore governed by characteristics of the items, in particular by the level and distribution of difficulty within the item set. Traditional approaches to test scoring and scaling are thus based on procedures involving reference to particular groups or item sets, and the interpretation of scores requires knowledge of the characteristics of these.

It was suggested in Section 2.2.3, in connection with the item statistics, that for certain item types (e.g. multiple-choice, true-false) it is possible for item scores to be artificially inflated by chance success. Clearly, this is also true of

person scores, and methods which aim to 'correct' scores for guessing have therefore been devised. The technique most frequently described involves subtracting from each person's number-correct score a proportion of the incorrect answers given, this proportion being determined by the number of answer options available (see e.g. Guilford & Fruchter, 1978:453-4).

As in the case of the analogous procedure for 'correcting' item scores, however, this is not felt to offer an entirely satisfactory solution, since it assumes that every incorrect response is the result of random guessing (Ebel, 1972:250), an assumption which seems unlikely to be justified. Ebel (1972:250) further notes that an adjustment of this type does nothing to ensure that a lucky guesser fares no better than an unlucky guesser; indeed, for the same number of guesses, the latter will be penalised more heavily, since more of the guesses he/she makes will appear as wrong answers. More serious, according to Guilford and Fruchter (1978:455), however, is the fact that the answer options in multiple-choice items are rarely equally attractive or plausible, so that guessing may actually involve a random choice from only 2 or 3 options, rather than from 4 or 5, resulting in undercorrection when the scoring formula is applied. It is primarily for this reason that they consider it preferable to employ instead a method for weighting right and wrong answers on the basis of correlations with performance on some relevant external criterion.

2.2.5 Requirements For Samples

The dependence of traditional test statistics on the samples used in their calculation, and hence the need for representative samples, has already been mentioned. A further important consideration, however, is that of sample size.

As Nunnally (1978:11,119) points out, psychometric theory is for the most part a large-sample theory, i.e. it assumes that large numbers of testees are used in test development and validation procedures. As an indication of the size of group that might be considered adequate, Thorndike (1982a:11) describes samples in the hundreds rather than in the thousands as being "of modest size", and Spearritt (1982:241) considers a sample size of 500 to be the minimum desirable for purposes of item analysis. The need for such large samples is perhaps not always appreciated by those engaged in the development of language tests.

2.2.6 Appraisal

Classical test theory, and the scoring, scaling and item analysis procedures that are usually associated with it, are generally seen to have contributed greatly to an increased understanding of psychological and educational testing, and to an improvement in the measures developed. Gulliksen (1950:1), for example, writes: "Since 1900 great progress has been made towards a unified quantitative theory that describes the behavior of test items and test scores under various conditions", and Lord (1974:107) describes classical theory as being: "... of great practical value in the design, construction, pretesting, scoring, statistical analysis, and interpretation of conventional tests of all kinds."

It is, however, also widely acknowledged that classical measurement theory has been unable to provide solutions to certain testing problems. Thorndike (1982a:5), for example, mentions the problem of deciding which reliability coefficient correctly indicates the proportion of true score variance when more than one such coefficient has been obtained using different procedures (e.g. test-retest, parallel forms, split-halves).

The restrictions of traditional item statistics were described in Section 2.2.3, and the deficiencies of traditional score scales are summarised by Wright (1968:86) as follows:

> "They have no zero point and no regular unit. Their meaning and estimated quality depend upon the specific set of items actually standardized and the particular ability distribution of the children who happened to appear in the standardizing sample."

Thus although, as Bejar (1983a:29) observes, the classical model "... has been the psychometric backbone of achievement testing over the last several decades", certain deficiencies remain. The section which follows is concerned with more recent developments in test theory which are seen as offering solutions to some of these problems.

2.3 ITEM RESPONSE THEORY

The rest of this chapter is concerned with the approach to test theory known variously as 'latent trait (LT) theory', 'item characteristic curve (ICC) theory', 'item response theory' (IRT), and, occasionally, as 'modern test theory' (Gustafsson, 1977:1; Hambleton, Swaminathan, Cook, Eignor & Gifford, 1978:468).

The first of these terms recalls the origins of this approach in psychology, where, as Lord and Novick (1968:359) explain, 'latent trait' denotes a psychological

dimension necessary for the psychological description of individuals, i.e. a hypothetical construct, such as those mentioned at the beginning of the chapter, which is assumed to underlie observed behaviour (Samejima, 1983:159). In the context of testing, latent traits are conceived of as characteristics or attributes which account for consistencies in the individual's responses to items (Wainer & Messick, 1983:343).

Anastasi (1983:346) comments that the term 'latent traits' has sometimes been taken to refer to fixed, unchanging, causal entities; however, as Samejima (1983:159) points out, latent traits should not be thought of as fixed, since a trait such as 'achievement' is capable of change or improvement, e.g. as a result of instruction. Lord and Novick (1968:358) further note that a trait orientation to psychological theory carries no necessary implication that traits exist in any physical or physiological sense. These comments apply equally to the notion of latent trait as it is used in the measurement context.

The term 'item characteristic curve theory', as will be seen later, derives from one of the concepts central to this approach, while 'modern test theory' emphasises the departure from the classical approach. The term 'item response theory', which will be used throughout here, is found in Frederic Lord's more recent work on the subject, and appears generally to be gaining currency. Samejima (1983:159) attributes the use of 'IRT' in preference to 'LT theory' to an effort on the part of some researchers to avoid the possible misinterpretation of 'trait', while Weiss (1983:2) views 'IRT' as emphasising the role both of test items and of testees' responses.

2.3.1 Central Concepts in IRT

The essential feature of an IRT approach is that a relationship is specified between observable performance on test items and the unobservable characteristics or abilities assumed to underlie this performance (Hambleton et al., 1978:469). The characteristic measured by a given set of items, whether a psychological attribute, a skill, or some aspect of educational achievement, is conceived of as an underlying continuum, often referred to as a latent trait or latent variable. Although the trait is usually viewed as being continuously distributed, no specific form of distribution (such as a normal distribution) needs to be assumed (Hulin et al., 1983:15).

This underlying continuum is represented by a numerical scale, upon which a person's standing can be estimated using his/her responses to suitable test items

(Hulin et al., 1983:15). Items measuring the trait are seen as being located on the same scale, according to the trait level they require of testees.

2.3.1.1 Person Ability and Item Difficulty

A person's standing on the scale is frequently called his/her 'ability'. As the use of this term is a potential source of misunderstanding, it must be emphasised that it refers simply to whatever characteristic, skill or area of understanding the test measures. As Rentz and Bashaw warn:

> "The term "ability" should not be mysterious: it should not be entrusted with any surplus meaning nor should it be regarded as a personal characteristic that is innate, inevitable or immutable. Use of the word "ability" is merely a convenience." (Rentz & Bashaw, 1977:182)

Lord (1974a:108), too, emphasises that the term 'ability' is used simply to mean the typical or expected performance of an individual in the area represented by the class of test questions.

An item's location on the scale is usually called its 'difficulty', particularly in the case of educational tests, which will be the main concern here. (Clearly, the concept of item difficulty is less applicable to the measurement of attitudes or personality traits, and location on the scale in this context is more appropriately thought of as the trait level embodied in the item.)

Central to IRT, therefore, is the notion that persons can be placed on a scale on the basis of their ability in a given area, and that items measuring this ability can be placed on the same scale. Thus there is "... a single scale ... which measures (is) both difficulty and ability simultaneously" (Pollitt, 1979:58). It is via this scale that the connection between items and respondents, which Traub and Wolfe (1981:378) describe as the essence of IRT, can be made.

2.3.1.2 Item Response Function/Item Characteristic Curve

The probability of a person making a particular kind of response to an item (correct or incorrect, in the case of dichotomously-scored items) is seen as being governed by the position of each on the common scale, i.e. by the person's ability and the item's difficulty, and sometimes also by additional properties of the item, such as its power to discriminate. Accordingly, the core of an IRT approach is a mathematical statement, sometimes referred to as an 'item response function' (see e.g. Lord, 1980:12; Swaminathan, 1983:24), which relates the probability of a correct response to these person and item parameters.

As will be seen in Section 2.3.3, this relationship is formulated in different ways by the various IRT models, usually known as 'item response models', which have been proposed. However, all take the general approach that the probability of a correct answer depends only on information about the person and the item. As Traub and Wolfe (1981:378) explain, in the case of unidimensional models, i.e. those concerned with the measurement of a single ability (a restriction which, they note, is almost always made in practice), the information about a person consists of just one numerical value (ability level), while the information about an item consists of one, two or three values, depending on the particular model adopted. They further note that the functional connection which an item response model specifies between probability of a correct response and information from these two sources is assumed to be appropriate for all persons in a given group responding to all items measuring the same trait. Thus given the necessary person and item parameters, the response of any person to any item can be predicted via the model of person-item interaction embodied in the item response function.

When the probability of a correct answer as given by the item response function is expressed for a particular item as a function of ability, this expression is referred to as the 'item characteristic curve' (ICC). When plotted, this provides a graphical representation of the way in which the probability of success on an item depends on ability level. An example of an ICC (adapted from Woods & Baker, 1985:124) is shown in Figure 2.1.

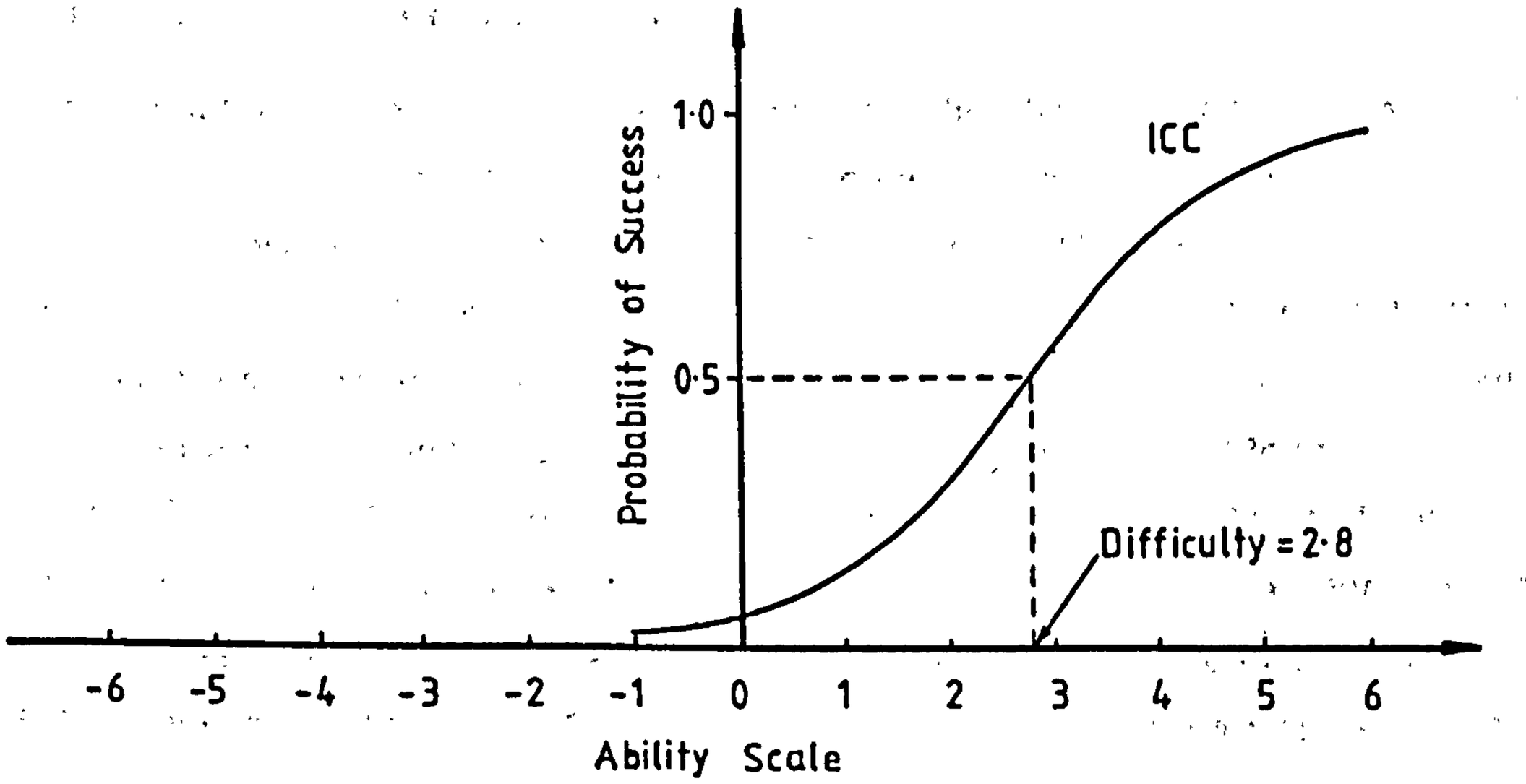


Figure 2.1 Example of an Item Characteristic Curve

A way of locating an item on the scale is to define item difficulty as being

equal to the ability scale value for which the probability of answering the item correctly is 0.5 (Traub & Wolfe, 1981:378); following this procedure, the difficulty value of the item depicted in Figure 2.1 would be approximately 2.8. The ICCs for items of different difficulty levels would differ in their positions with respect to the ability scale axis: the ICC of an item easier than the one shown above would appear further to the left, while that of a more difficult item would appear further to the right.

Hulin et al. (1983:19-20) make a distinction between theoretical ICCs, i.e. the mathematical form of the ICC as specified by an item response model, and empirical ICCs, which are obtained from a set of response data by determining the proportions of testees at various levels of ability who answered a given item correctly. Empirical ICCs have in some cases been used as the starting point for the development of item response models; however, even when the theoretical form of the ICC is selected for other reasons, empirical ICCs can be used to provide a check on whether or not it satisfactorily models the performance of particular groups of persons on particular sets of items.

Clearly, before the suitability of the chosen item response function can be assessed, or any practical applications undertaken, estimates of the person and item parameters must be obtained from a set of response data. These values are estimated by means of equations deriving from the response model itself. Some methods of estimation will be outlined in Section 2.3.5.

2.3.2 Development and Current Impact

The beginnings of an IRT approach can be traced back to work in psychophysics at the turn of the century (Hambleton & van der Linden, 1982:373), and are also evident in studies of growth and mortality rates undertaken by biometricians in the 1920s (Wright & Stone, 1979:ix). The earliest applications of IRT ideas in psychometrics are found in the work of Thurstone in the 1920s, and that of Ferguson, Lawley, Tucker and Guttman in the 1940s (Bejar, 1983a:29; Wainer, 1983:xvi). The contribution of Lazarsfeld in the 1950s is frequently acknowledged (see e.g. Gustafsson, 1977:1; Hambleton & van der Linden, 1982:373; Bejar, 1983a:29); however, the most important developmental work in IRT for application to tests of ability and achievement was carried out in the 1950s and 60s by Frederic Lord and Allan Birnbaum in the U.S.A., and, in an independent but parallel development in Europe, by Georg Rasch (Weiss, 1983:3-4). The relevant theoretical background is set out in Lord and Novick's (1968) *Statistical Theories of Mental Test Scores*, which includes four chapters

by Birnbaum, and in Rasch's (1960) *Probabilistic Models for some Intelligence and Attainment Tests*.

Hambleton et al. (1978:468) attribute the slow progress in the implementation of IRT at that time to the mathematical complexity of the field, the shortage of convenient and efficient computer programs, and scepticism as to the benefits to be gained. Weiss (1983:4) notes that except for some of the procedures deriving from Rasch's work, implementation was not feasible until suitable computing facilities became available in the late 1960s. However, as the computational problems began to be solved, and successful applications published, IRT became an area of great interest to measurement specialists (Hambleton et al., 1978:468), with further influential contributions to the application and/or extension of IRT being made by, for example, Benjamin Wright, Gerhard Fischer, R. Darrell Bock and Fumiko Samejima (Hambleton & van der Linden, 1982:373; Bejar, 1983a:29).

Baker (1977:151-2) comments that the delay between the development of IRT and its use in practical item analysis work seems to have been unusually long; the reasons he suggests for this are the "rather sophisticated level of mathematics" used in presenting the developments, and the paucity of articles explaining them to practitioners. A further contributory factor, according to Hambleton and Cook (1977:76) was the failure of some mental test data to satisfy the strong assumptions upon which IRT is based.

Although it is still the case that much of the theoretical work on IRT is not easily accessible to those who are not mathematicians or statisticians, there are a number of relatively non-mathematical introductions to the area, such as those by Willmott and Fowles (1974), Baker (1977), Hambleton and Cook (1977), Hambleton et al. (1978), Traub and Wolfe (1981), Hulin et al. (1983) and Hambleton and Swaminathan (1985). Furthermore, several computer programs for the application of IRT procedures are now available, including CALFIT (Wright & Mead, 1975), BICAL (Wright, Mead & Bell, 1980), MICROSCALE (Wright & Linacre, 1984), PML (Gustafsson, 1981), DISLOC (Andrich, De'Ath & Lyne, 1982) and LOGIST (Wingersky, Barton & Lord, 1982).

This increased accessibility, coupled with the publication of work demonstrating the potential usefulness of IRT (see, for example, the special issue of Journal of Educational Measurement, Summer 1977), has led to a surge of interest among test developers and researchers, so that by the early 1980s Bejar was able to write: "There are indications that the theory is now reaching the practitioner and may in fact prove to be the "standard" psychometric model"

(1983a:29-30). Indeed, Hambleton and Murray noted in 1983 that IRT was already being used by almost all of the major test publishers in the U.S.A., as well as by many state departments of education and industrial and professional organisations (p.71). In the same year, Weiss (1983:7) described IRT as an area which promised to have "... profound implications for the improvement of psychological measurement and for the solution of a variety of applied problems that have not been adequately solved by over a half century of classical psychometrics."

Interest in the use of IRT in the field of second/foreign language testing has also begun to grow in recent years, as is indicated by the increasing number of IRT-related presentations at professional meetings, and the appearance in the language testing literature of the first papers concerned with IRT (see, for example, de Jong, 1983; Perkins & Miller, 1984; Henning, 1984; Woods & Baker, 1985; Griffin, 1985; Stansfield, 1986; Pollitt & Hutchinson, 1987). Indeed, a recent language testing handbook (Henning, 1987) includes sections on IRT.

2.3.3 The IRT Family of Models

So far, item response theory has been presented in general terms, as a class of mathematical models developed for use in measuring individuals' ability from their responses to test items. In this section, a brief account is given of the types of model usually viewed as belonging to the IRT family, and of the three most widely-discussed models pertaining to dichotomously-scored, unidimensional test data.

2.3.3.1 Model Types

Comprehensive accounts of the various types of model within the IRT class are provided e.g. by Hambleton and Cook (1977), Hulin et al. (1983) and Hambleton and Swaminathan (1985). A classification scheme for these models in terms of response level (no. of response categories for each item), parametric structure (no. of item parameters defined) and statistical assumptions (form of the ICC) is set out by Bejar (1983a:31-33;1983b:10-11).

A fourth classification variable mentioned by Bejar is that of the dimensionality of the latent space, i.e. whether the model is intended for use with unidimensional or multidimensional data sets. As Hambleton and Cook (1977) explain, the dimensionality of the data depends on the number of traits underlying performance on the items. They note that models which allow for

more than one trait are complex, and less well-developed than unidimensional models.

Within the sub-class of unidimensional models, models may be differentiated according to the type of response data to which they apply. This study is concerned with dichotomously-scored response data; there are, however, also models for polytomous and continuous response data (Bejar, 1983a). Within the polytomous class, Samejima (1983) distinguishes cases in which items are scored into more than two graded response categories, and those of the 'nominal response' type, in which the response categories are not explicitly ordered. Thus as regards models which would be applicable in the context of educational testing, there are, in addition to those for use with items scored correct or incorrect, extensions to these which are appropriate for data from rating scales, and for items scored according to a 'partial credit' approach (see, for example, Wright and Masters, 1982). Models for the dichotomous case have also been adapted so that account may be taken of omitted responses (see e.g. Lord, 1974b).

Within the class of models for dichotomously-scored, unidimensional test data, there are differences in the nature of the relationship specified between success on items and person and item characteristics. A simple response model such as Guttman's 'perfect scale' model, for example, would state that a person's probability of success on an item is either 0 or 1, depending on whether the item is above or below the person's ability level. The relationship specified in this case is a deterministic one, allowing for no source of error at any stage in the testing process. Clearly, this is an unrealistic expectation, since, as Samejima (1983:159) points out, a large number of factors may contribute to the individual's eventual response to an item. Thus it is generally agreed (see e.g. Rasch, 1960; Lord, 1974a; Pollitt, 1979; Samejima, 1983) that it is preferable, indeed necessary, to formulate the model in terms of a probabilistic rather than a deterministic relationship.

The various probabilistic models which have been formulated differ in the mathematical form (and hence the shape) of their item characteristic curves. As Hambleton and Cook (1977) illustrate, both linear and non-linear ICCs have been adopted. In one of the models they describe, the latent linear model, the ICCs take the form of straight lines which vary in their intercepts (points at which they meet the x-axis) and in their slopes. The intercept of the ICC with the x-axis indicates the difficulty level of the item, and the gradient of the line represents

its discriminating power: the sharp increase in the probability of success on a highly discriminating item as ability increases will be reflected in a steep ICC. On a less discriminating item, on the other hand, an equivalent increase in the probability of success will require a larger increase in ability, and thus the ICC for such an item will be less steep.

More commonly, however, the ICCs specified in item response models take the form of elongated 'S'-shaped curves such as that shown in Figure 2.1. Much of Lord's early work (see e.g. Lord, 1952) was concerned with ICCs in the form of normal ogives, i.e. with the shape of a cumulative normal distribution. Use of the normal ogive, however, proved to be difficult from a mathematical point of view, and so the normal ogive model has largely been superseded by logistic models, i.e. models in which the ICCs take the form of logistic functions. As Birnbaum (1968) explains, the logistic ogive is very similar to the normal ogive, but has the advantage of being mathematically more convenient to use.

The models which now receive the most attention in the IRT literature are thus the logistic response models. Three such models are described in the next section.

2.3.3.2 Three Major Logistic Response Models

The three response models outlined here are frequently presented as the major competing models for unidimensional, dichotomously-scored test data. As will be seen later, this view is not shared by those who perceive important differences in the approaches to measurement which they imply. However, in terms of their mathematical form, at least, these models may be regarded as members of the same sub-class, and it is therefore convenient to describe them together.

As has already been indicated, a common feature of these models is that the ICCs take the form of logistic curves. The difference between them, however, is in the number of variables, or parameters, required to describe an item. This difference is reflected in the names by which two of these models are usually known (the 'two-parameter logistic model' and the 'three-parameter logistic model'). The third, however, though sometimes referred to as the 'one-parameter logistic model', is more commonly known as the Rasch model, after the mathematician who developed it. (The development of the two other models is attributed largely to Lord and Birnbaum.)

It should be noted that the numbers in these names refer to the number of item parameters. All of these models, of course, also involve a person (ability) parameter, and thus the Rasch model, for example, involves a total of two parameters. This occasionally appears to cause confusion; here, however, the convention of referring to models by the number of item parameters is followed.

The two-parameter logistic model has parameters for item difficulty and item discriminating power. As in the case of the linear model mentioned in Section 2.3.3.1, the steepness of the slope of the ICC represents the item's power to discriminate, though in this case the ICC is an elongated 'S'-shaped curve rather than a straight line. The item's difficulty value is defined as the point on the ability scale at which the slope of the ICC is at a maximum (Hambleton & Cook, 1977), i.e. at the midpoint of the ICC. Below this point, probability of success on the item tends to zero as ability decreases, and above it, probability of success tends to 1 as ability increases.

The fact that differences in discriminating power, and hence in the slopes of ICCs, are allowed for in this model means that the ICCs, although members of the same general family of curves (logistic), may differ somewhat in shape. As can be seen from the examples in Figure 2.2, they may also cross. A further feature of these curves is that they approach the x-axis at the lower extreme of ability: this reflects the assumption that testees do not answer correctly by random guessing. In a test which allowed correct guessing, the probability of success even for a person placed infinitely low on the ability scale would be further above zero than shown in the example ICCs for this model.

The three-parameter logistic model attempts to account for the possibility that correct guessing might occur. As far as the difficulty and discrimination parameters are concerned, this model is the same as that described above. The difference between the two- and three-parameter models is that in the latter, probability of success does not tend to zero as ability decreases. This is shown in the example curves in Figure 2.2: the lower asymptotes of the ICCs in the three-parameter model do not approach the x-axis as they do in the case of the two-parameter model. The third parameter, which distinguishes these two models, is thus a 'guessing' or 'pseudo-chance level' parameter, defined by Lord (1980) as the probability that a person completely lacking ability (i.e. with no knowledge or skill in the tested area) will succeed on the item.

The Rasch model is similar to the two-parameter model in that probability of success is assumed not to be affected by the possibility of guessing. It differs

from both other models, however, in that the ICCs can differ only in their translation along the ability scale, and not in their shape. Thus while item difficulty varies, discriminating power is assumed to be the same for all items. It is therefore not possible for the ICCs to cross as they do in the two other models. Example curves for the Rasch model are also shown in Figure 2.2. (Examples of ICCs for these, and other, response models may be found in e.g. Hambleton & Cook, 1977 and Hambleton & Swaminathan, 1985).

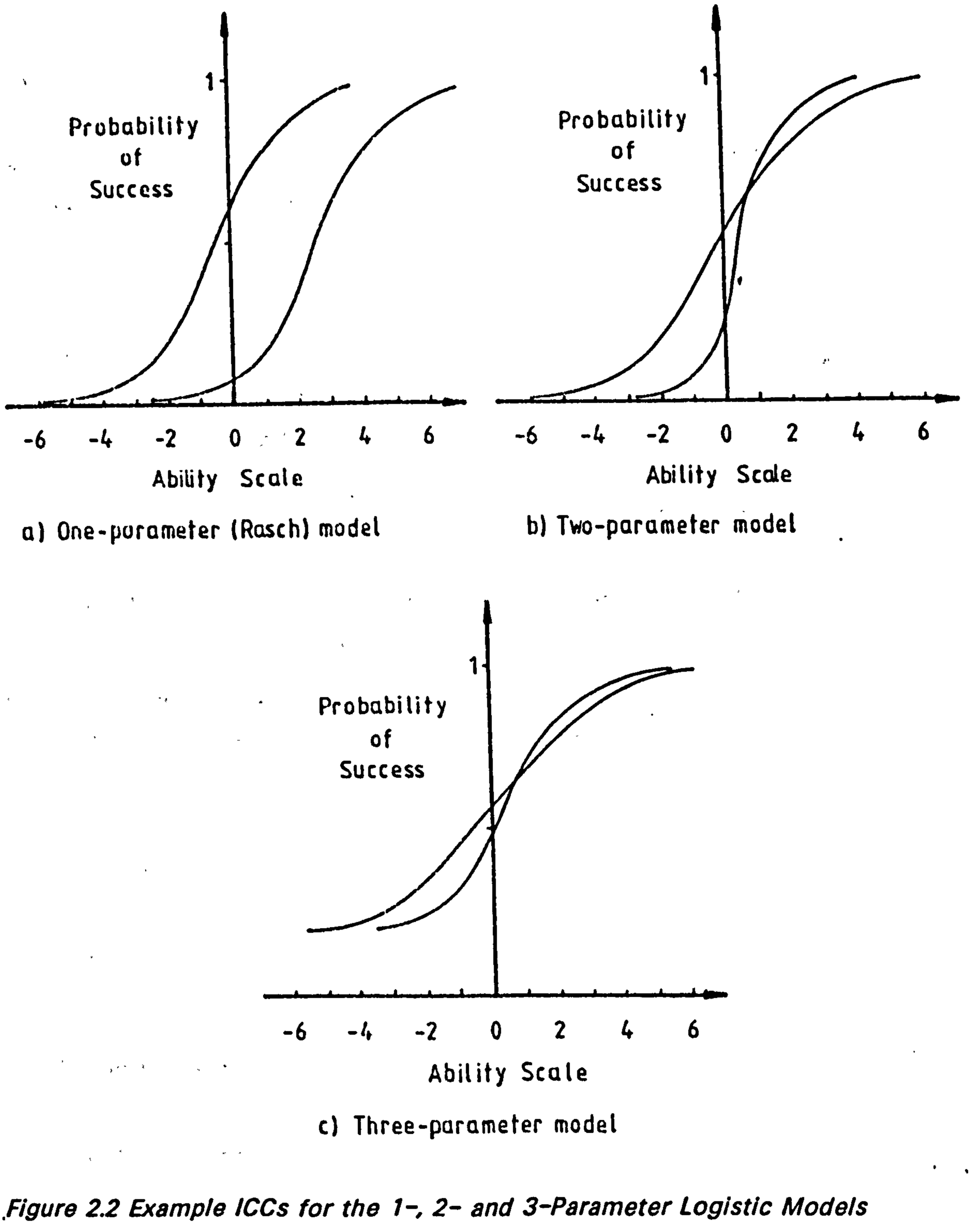


Figure 2.2 Example ICCs for the 1-, 2- and 3-Parameter Logistic Models

2.3.3.3 Mathematical Forms of the 1-, 2- and 3-Parameter Logistic Models

The mathematical forms of the three models described in the previous section are set out below. It should be noted that each model can be written in various ways, and that different authors use different notation. The forms of the equations presented here are those given by Gustafsson (1977): these have been selected on grounds of simplicity, and in order to facilitate direct comparison of

the models.

For the sake of consistency with later sections, however, different notation from that of Gustafsson is used here. Although most authors use Greek symbols to denote parameters (as opposed to estimates of parameters), observation of this convention does not seem necessary for the purposes of this section, and thus the notation used is as follows:

b = the ability parameter of a randomly selected person;

d_i = the difficulty parameter of item i;

a_i = the discrimination parameter of item i;

c_i = the lower asymptote of the ICC, often called the 'guessing' or 'pseudo-chance level' parameter of item i;

$p_i(b)$ = the probability that a person with ability b will answer item i correctly;

$\exp(x)$ = the natural constant e (approximate value 2.71828) raised to the power x.

The way in which these person and item parameters enter into the probabilistic relationships specified by the three models can be seen from the equations below.

(i) ICC in the one-parameter (Rasch) model:

$$p_i(b) = \frac{\exp(b-d_i)}{1 + \exp(b-d_i)}$$

(ii) ICC in the two-parameter model:

$$p_i(b) = \frac{\exp[a_i(b-d_i)]}{1 + \exp[a_i(b-d_i)]}$$

(iii) ICC in the three-parameter model:

$$p_i(b) = c_i + (1-c_i)\frac{\exp[a_i(b-d_i)]}{1 + \exp[a_i(b-d_i)]}$$

2.3.4 IRT Assumptions

The main assumptions made in these response models are those relating to the form of the item characteristic curve, test unidimensionality, and local statistical independence (see e.g. Gustafsson, 1977:9; Traub & Wolfe, 1981:387). A brief explanation of each of these is given below.

2.3.4.1 Form of the ICC

The item characteristic curve is a mathematical function that specifies the way in which a person's responses depend upon his/her ability. More precisely, it states the relationship between probability of a correct answer to an item and trait (ability) level. As Gustafsson (1977:12) explains, the form of the ICC to be used must be decided upon in order for it to be possible to formulate the statistical models from which the equations for parameter estimation are determined. (An exception to this, noted by Hambleton and Cook (1977:80), is the work of Lord (1970), in which the form of ICCs is not specified in advance, but chosen so as to maximise their fit to the data.)

It has been seen that the various response models differ in respect of the ICCs with which they operate. However, the particular ICC adopted is assumed to provide a plausible representation of the relationship between performance on test items and ability: the assumption is thus made that the mathematical form of the ICC is correct for the data set in question.

2.3.4.2 Unidimensionality

A second major assumption made in these models is that the item set is unidimensional, i.e. that the items measure a single ability or trait (Hambleton & Cook, 1977:77). This is not to say that the attribute underlying performance on a test needs to be psychologically simple, but rather that it should, whether simple or complex, be approximately the same for all the items in the test (Thorndike,1982a:9). Lumsden (1976:267) suggests that there has been a tendency to confuse unidimensionality with theoretical singularity, and points out that a test can be unidimensional even though the ability it measures might have to be viewed not as a single theoretical construct, but as a compound with constructs as elements. He writes: "A unidimensional test does have a single attribute but the attribute is complex." A similar view is expressed by Bejar (1983a:31):

> "It should be pointed out that unidimensionality does not imply that performance on the items is due to a single psychological

> process. In fact, a variety of psychological processes are involved in the act of responding to a set of items. However, as long as they function in unison - that is, the performance on each item is affected by the same processes and in the same form - unidimensionality will hold."

The assumption of unidimensionality in IRT thus requires that the item set be relatively homogeneous; indeed, it is frequently argued (see e.g. Lord, 1974a:108; Lumsden, 1976:266; Gustafsson, 1977:9) that meaningful measurement under any approach is possible only if the test is unidimensional in this sense.

A statistical definition of unidimensionality as it relates to IRT, based on Lord and Novick's (1968:359) general definition of dimensionality of any order, is set out e.g. by Gustafsson (1977:9; 1980:207) and Hambleton and Swaminathan (1985:18-20). This states that for a test to be unidimensional, the distributions of scores for persons at any specified ability level must be identical for the various subpopulations within the population for whom the test is intended. Hambleton and Swaminathan (1985:18-19) note that if these conditional distributions vary across subpopulations, the test must be measuring something other than the single ability of interest, and is therefore not unidimensional for that population. To illustrate the point that a test might be unidimensional for one population but not for another, they cite the example of a mathematics test in which a certain level of reading comprehension is required in order to understand the questions: for a subpopulation with sufficient reading ability, performance will depend on maths ability alone, whereas for a subpopulation in which not all persons can understand the questions, performance will be affected by both maths ability and reading ability.

Implicit in the unidimensionality assumption as outlined here, therefore, is the requirement that performance on a test should not be influenced to any great extent by factors such as the effects of a time limit (Hambleton & Swaminathan, 1985:30). Clearly, where conditions are such that testees are prevented by lack of time from attempting all the items, the test can no longer be considered as a measure of a single ability even if the items are homogeneous with respect to content, since an additional, confounding variable will have been introduced.

2.3.4.3 Local Independence

The third main assumption in these models is that of local independence. Again, the original definition given by Lord and Novick (1968:361) is presented, often in a somewhat simplified form, by many other authors, including Gustafsson (1977:11), Hambleton and Cook (1977:77), Traub and Wolfe (1981:387),

Thorndike (1982b:82), Hulin et al. (1983:41-43) and Hambleton and Swaminathan (1985:22-24).

The principle of local independence states that for persons located at any given point on the ability scale, the probability of a person answering any one item correctly is not affected by information regarding that person's success or failure on any other item(s) (Lord, 1974a:110; Thorndike, 1982b:82). Hulin et al. (1983:42) note that stated more generally, this means that (given, of course, the relevant item parameters) all the information concerning the probability of a correct or incorrect response is contained in the ability parameter, and that if this parameter is known, then observing a person's responses to one or more of the items in a test provides no additional information about his/her responses to any other(s).

As Gustafsson (1977:11) observes, the assumption of local independence implies that a person's answer on one item does not influence his/her answer to any other; Hambleton and Swaminathan (1985:23) point out that the content of one item must not, therefore, provide any clues to the answer to another, and that e.g. the order in which the items are administered must not affect performance. Expressed in statistical terms, the local independence assumption requires that a person's probability of obtaining a particular pattern of correct and incorrect answers be equal to the product of the probabilities for each individual answer (Gustafsson, 1977:11; Hambleton & Cook, 1977:77; Hambleton & Swaminathan, 1985:23). Lord and Novick (1968:361) note that this is an automatic consequence of the proper choice of latent variables underlying performance on a test: for the particular case with which we are concerned here, i.e. that of unidimensional tests, it can therefore be stated that the assumption of local independence will necessarily be satisfied if the test measures a single ability (Hambleton, 1979:16). Indeed, the local independence assumption is sometimes said to be equivalent to the unidimensionality assumption (see e.g. Gustafsson, 1977:11; Hambleton & Swaminathan, 1985:22). Lord (1980:19) views it as following automatically from unidimensionality, rather than as an additional assumption.

Hulin et al. (1983:43) state the relationship between local independence and unidimensionality as being that each implies the other, a point which can be illustrated using the examples offered by Lord (1974a) and Hambleton and Swaminathan (1985). Lord (1974a:110) explains that in a case where local independence does not hold, and hence some people's probability of answering

certain items all correctly is, say, greater than the product of the individual probabilities for each correct answer, then those people can be expected to obtain systematically higher scores than others of the same ability level. This would indicate that the test measures more than one ability, i.e. that the unidimensionality assumption does not hold.

Hambleton and Swaminathan (1985:23) present the same argument, but from a different starting point: they note that where a test measures on e.g. two dimensions, people of a relatively high level in the second ability will have greater chances of success on the items which tap this ability than those of a lower level. Thus for a fixed ability level (taking the test as a whole), performance across certain items will be correlated, a clear violation of the local independence assumption.

Lord and Novick (1968:361) emphasise that the local independence assumption does not imply that item scores are uncorrelated for the whole group of persons: as Hambleton and Swaminathan (1985:24) point out, positive correlations between pairs of items will result whenever there is variation among persons with respect to the measured ability. They note that it does, however, imply that item scores are uncorrelated for a given ability level: thus where two items are linked by a specific trait, in addition to the trait which they share with the rest of the items in the test, local independence will not hold (Traub & Wolfe, 1981:387).

It is sometimes remarked that unidimensional item response models are based on strong (i.e. restrictive) assumptions (see e.g. Traub & Wolfe, 1981:387; Hambleton & Swaminathan, 1985:155). In Section 2.3.6, mention will be made of some possible ways of assessing the extent to which these assumptions might be met by sets of test data. First, however, an indication is given of the way in which the person and item parameters are estimated in these models, and some of the properties of these estimates are considered.

2.3.5 Ability and Difficulty Estimates in IRT

From item response models such as those set out in Section 2.3.3.3, equations can be derived for the estimation, for a given set of response data, of the person and item parameters. In the case of the Rasch model, which assumes that only person ability and item difficulty affect the probability of a correct response, estimates need to be obtained only for the ability of each person in the group and for the difficulty of each item in the set. Where additional item parameters are seen as influencing the outcome, as in the two- and

three-parameter models, estimates of these must also be obtained for each item; discussion here, however, is confined to the estimation of abilities and difficulties.

As Thorndike (1982a:10) explains, item response models share the view of a test item as providing an indication of the individual's standing on the ability/difficulty scale: a person who succeeds on an item is likely to be positioned above it on the scale, while a person who answers the same item incorrectly is likely to fall below it. Estimating a person's ability, which is the IRT equivalent of scoring a test, is therefore a matter of finding the point on the scale that best summarises his/her performance on a set of items. Similarly, estimating the difficulty of an item involves using the information contained in responses to it by a group of persons to find its likely location on the scale. In traditional terms, this is analogous to calculating the item's facility value; however, Traub and Wolfe (1981:406) point out that classical item statistics are normally used only in item selection, whereas IRT item statistics also enter into procedures for scoring, calibrating and equating tests. They therefore emphasise the need in IRT for precise and unbiased parameter estimates.

Since IRT depends for its application on the availability of practical methods for obtaining accurate parameter estimates, this aspect of IRT has received a great deal of attention in the literature, and considerable effort has been devoted to trying to solve some of the statistical and computational problems which it entails. As a result, a number of possible methods have been proposed. These differ in their complexity and accuracy, and in respect of the particular models to which they apply. For the Rasch model, in which item difficulty is the only item parameter specified, parameter estimation is less problematic than for the models involving additional item parameters: indeed, it is sometimes stated (see e.g. Gustafsson, 1977:15; Wright, 1977a:102) that the theoretical and practical problems associated with parameter estimation have been completely solved only for the Rasch model. However, Gustafsson (1977:17) acknowledges that although this is true of the estimation of the item parameters, the problem of obtaining unbiased ability estimates remains. Furthermore, for parameter estimation under the Rasch model alone, various approaches are possible (see, for example, the discussion by Gustafsson, 1980:209), and there are differing views as to which of these should be adopted, given that here, as so often in measurement, the desire for precision may conflict with the need for practicability.

2.3.5.1 Methods of Estimation

The main methods available for parameter estimation in IRT are outlined, and their relative merits discussed, by e.g. Traub and Wolfe (1981:406-413), Swaminathan (1983), Bejar (1983a:39-43) and Hulin et al. (1983:46-53). Those based on maximum likelihood estimation are the most widely used, and, according to Hulin et al. (1983:46), probably the most important from a theoretical point of view. Traub and Wolfe (1981:407) observe that maximum likelihood estimation represents a conventional statistical approach to the estimation of IRT parameters, and most authors refer to the desirable and useful properties, including consistency and efficiency, that maximum likelihood estimates have been shown to possess (see e.g. Traub & Wolfe, 1981:407; Hulin et al., 1983:48; Swaminathan, 1983:30).

As Hulin et al. (1983:46) note, maximum likelihood estimation rests upon the simple idea that the parameter estimates chosen should be the values which make the observed data set appear most likely in the light of the particular model being used. Expressed in statistical terms, this involves maximising the likelihood function for the observed response matrix. The likelihood function is defined as the product of the probabilities, as specified by the model, of all the correct and incorrect responses in the data set (Traub & Wolfe, 1981:407).

For an observed data set, the outcome of each person's attempt at each item is known, and is represented by 1 for a correct answer and 0 for an incorrect answer. Where the item parameters are also known (or, more realistically, where estimates of them are available from a previous calibration), obtaining maximum likelihood ability estimates is, notwithstanding certain numerical problems, considered to be a relatively straightforward matter (Swaminathan, 1983:32). It should be noted, however, that for persons with zero or perfect scores, and for items which have been answered either all correctly or all incorrectly, finite maximum likelihood estimates are not available (Hambleton & Swaminathan, 1985:86); thus the position of a person or an item on the scale cannot be estimated in this way unless the response data for the person or item in question contains at least one right and one wrong response.

As Bejar (1983a:40) explains, just as one speaks in IRT of the probability of a correct response to a single item, one can also speak of the probability of observing a particular response vector (the pattern of 1s and 0s for a person on a set of items). This probability is given by multiplying together the individual probabilities for all the responses in the vector; as Swaminathan (1983:27) points

out, it is more properly referred to as a likelihood function than as a probability, since the actual responses are already known. Just as the item characteristic curve expresses probability of success as a function of ability, the likelihood function expresses the likelihood of the response vector as a function of ability. The ability value for which this likelihood is a maximum is the maximum likelihood estimate of ability (Bejar, 1983a:40). Finding this value for each person, by solving the likelihood equations derived from the model, requires the use of numerical methods, and has to be done by computer. However, Bejar (1983a:39) notes that suitable computer programs need not be difficult to develop, and comments that the process of ability estimation is nevertheless "conceptually straightforward".

Of course, the estimation of ability when item parameters are known represents a somewhat unrealistic simplification of the estimation problem, since in most IRT applications it will be necessary to estimate both ability and item parameters from the same data set (Hulin et al., 1983:48,52). Estimation of this kind, which is often called 'joint estimation', presents a number of statistical problems, particularly for the two- and three-parameter models, under which each person in the sample introduces an additional person parameter, with the result that the maximum likelihood estimates for the item parameters may be inconsistent (Gustafsson, 1977:15; Traub & Wolfe, 1981:408).

A general approach to the joint estimation of IRT parameters, described by Birnbaum (1968) and Lord (1980), and summarised by Hulin et al. (1983:52), involves using approximate methods to obtain initial estimates of both sets of parameters, and then solving the likelihood equations for the item parameters while holding the initial ability estimates constant. The new item parameter estimates are then used in re-estimating the ability parameters, and the resulting new ability estimates in re-estimating the item parameters, in an iterative process, until both sets of estimates converge (i.e. until the differences between successive re-estimations are negligibly small).

Although, as is frequently pointed out, the properties of maximum likelihood estimates obtained in this way are unknown, Hulin et al. (1983:53) consider it likely that the results will be sufficiently accurate for many applications of IRT, and Traub and Wolfe (1981:412) observe that there is "a considerable body of experience" which suggests that programs based on joint estimation methods "... usually work and produce useful output."

As Swaminathan (1983:35) notes, joint estimation is considerably less

problematic when the Rasch model is used, because in this case the number-correct score is a sufficient statistic for ability estimation, i.e. it contains all the necessary information from the response pattern (Traub & Wolfe, 1981:408-9), and a person's ability estimate is a function of his/her raw score. Thus instead of having to calculate a separate ability estimate for each person in the sample, it is necessary only to obtain one for each raw score group: persons gaining the same raw score are considered to have the same ability level.

The availability of this sufficient statistic for ability estimation in the one-parameter model makes possible a further method of maximum likelihood estimation, usually known as 'conditional maximum likelihood estimation', which is described by Traub and Wolfe (1981:408) as offering a practical and exact solution to the problem of estimating item parameters. Unlike methods such as the iterative procedure outlined above, in which person and item parameters are estimated simultaneously (an approach sometimes called 'unconditional' estimation), the conditional maximum likelihood method involves expressing the likelihood function for estimating the item parameters in the item parameters only, so that although it contains functions of the number of persons in each raw score group, it is free of the individual ability values, and yields unbiased item parameter estimates (Gustafsson, 1980:209; Traub & Wolfe, 1981:408).

Gustafsson (1980:210) considers this method to be theoretically superior to unconditional estimation methods, and claims that the computational problems involved have now been overcome for tests of up to 100 items. However, estimates obtained using unconditional methods can be corrected to make them very similar to conditional maximum likelihood estimates (Wright & Douglas, 1977), and unconditional estimation continues to be widely used in computer programs for the application of IRT (see, for example, BICAL (Wright, Mead & Bell, 1980)).

Since conditional maximum likelihood estimation procedures cannot be applied in the case of the two- and three-parameter models, an alternative approach, known as 'marginal maximum likelihood estimation' has been proposed. However, Traub and Wolfe (1981:40) and Swaminathan (1983:38) indicate that this method is not yet sufficiently developed for general use.

It is important to note that the terms 'conditional' and 'unconditional' estimation have been used by some authors, notably Bock and Lieberman (1970), with almost the opposite meanings to those outlined here (Gustafsson, 1977:16). As Subkoviak and Baker (1977:304) explain, in this other usage, 'conditional'

estimation denotes a procedure in which one set of parameters is known, and the other estimated: it is therefore applied also to the simultaneous (joint) estimation of sets of parameters (Gustafsson, 1977:16). 'Unconditional' estimation, on the other hand, denotes a procedure in which one set of parameters is removed mathematically from the estimation process for the other set. Although Gustafsson (1977:16) considers that this second use of the terms 'conditional' and 'unconditional' deviates from their usual use in mathematical statistics, it has been adopted in some of the recent literature, with the result that discussions of estimation procedures may be confusing unless terms are clearly defined. These terms will be used here according to the original definitions given.

In addition to maximum likelihood estimation, and the various approximate methods suggested in the literature, another type of estimation, based on Bayesian procedures, has received some attention. In this approach, prior information about the parameter is incorporated into the estimation procedure. Although it is thought that this may result in estimates which are more accurate or meaningful than those obtained by other methods (Hambleton et al., 1978:485; Swaminathan, 1983:31), the use of such procedures is possible only when prior information about the distribution or levels of the parameter (e.g. the testees' ability levels) is available, or, as Bejar (1983a:41) observes, when the tester is willing to make assumptions about these.

A general point concerning parameter estimation, made by Traub and Wolfe (1981:409), is that in practical applications item calibration should be carried out on large person samples, so that the item parameter estimates will be quite close to the true values. As Rasch (1960,1980) points out, item difficulties will in any case usually be estimated more accurately than person abilities, since in practice the number of items that can be administered to a group is limited, while the number of persons who can be tested on a set of items need not be restricted: a set of response data will thus frequently contain more information about each item in the set than about each person in the group.

2.3.5.2 Information and Precision of Estimates

The notion of 'information', as formalised by Birnbaum (1968), is of great importance in IRT, since it provides the basis for assessing the degree of precision with which each person and item parameter has been estimated, and hence for determining the degree of confidence with which the estimates may be used.

Following Birnbaum (1968:454), the information about a person's ability level contained in his/her response to a given item is defined by means of an 'item information function'. In its general form, this function is written in terms of the item characteristic curve and the person's probability of success on the item multiplied by his/her probability of failure. The precise form of the item information function will depend on the particular item response model being used: in the case of the Rasch model, for example, this expression reduces to the product of the probability of a correct answer and the probability of an incorrect answer.

The information about a person's ability level contained in his/her responses to a whole set of items is found from the 'test information function', which Birnbaum (1968) defines as the sum of all the individual item information functions for the person. Thus as Lord (1968:1008) notes, an important property of information as defined in this way is that it is additive. The notion of information is therefore of use in selecting sets of items which will result in the most accurate measurement at given levels of ability.

A function of the same form as the item information function can be written to define the information about an item's difficulty level contained in the response of a person at a given ability level, and these too can be summed for each item across the group of persons to quantify the information about each item's difficulty contained in the whole set of responses.

Although the value of the information function with respect to a parameter can itself be used as a measure of the precision with which the parameter has been estimated, the index more frequently used for this purpose is the standard error of estimation. As Samejima (1977:236) explains, when a maximum likelihood estimation procedure has been used, the standard error associated with each estimate is given by the reciprocal of the square root of the information value.

Because of the way in which the model probabilities for correct and incorrect responses enter into the item information function, information is greatest (and hence the standard error lowest) when the person has equal chances of succeeding or failing on the item, i.e. when the person's ability is equal to, or, in the case of the three-parameter model, only slightly higher than, the item's difficulty. Thus generally speaking, the most precise estimates of the person and item parameters will be obtained when the persons and items are well-matched in levels.

2.3.5.3 Properties of the Parameter Estimates

It is useful at this point to make explicit the distinction between parameters and parameter estimates. The parameters, which appear in the mathematical statement of an item response model, represent the properties of persons and items that are considered to affect the outcome when persons attempt items. The parameter estimates are the numerical values obtained for each person and item when a set of response data is analysed in accordance with an item response model. These estimates are given by a statistical estimation procedure in which equations deriving from the model are solved.

Since item difficulty in IRT is seen as representing an intrinsic property of items, it is conceived of as being independent of the abilities of any particular group of testees. Similarly, person ability is seen not as being indexed to a test (as defined by a particular set of items), but as representing persons' levels on the measured trait. This is the conceptual basis of the parameter 'invariance' feature which is one of the fundamental characteristics of IRT, and on which the possibility of achieving 'sample-free' or 'subpopulation-independent' item calibration and 'item-free' or 'test-free' person measurement depends.

The sense in which IRT item parameters can be said to be invariant is best explained via a consideration of the properties of item characteristic curves. It was noted in Section 2.3.1.2 that an ICC expresses the probability of a correct answer to a given item as a function of ability. As Hambleton and Cook (1977:80) point out, the probability that an individual of a given ability level will answer the item correctly does not depend on how many other people are located at the same point, or indeed at a different point, on the ability continuum. Hence the distribution of ability within a particular sample has no influence on the shape of the ICC, or on its position with respect to the ability scale axis. Nor is the ICC affected by the characteristics of other items appearing in the test. Thus, as Rudner (1983:952) observes, even if the item is transplanted into a different test of the same trait, or administered to a different sample of testees, the ICC will be the same. Graphical illustrations of the invariance of ICCs across samples are provided by Baker (1977:170), Hambleton (1980:77) and Hulin et al. (1983:44). Hulin et al. note that provided the form of the ICCs is correct for some population of testees, and provided the probability of a correct answer is a function of the single ability represented by the ability scale, then subpopulations formed in any manner will share a common ICC.

As regards the actual estimation of ability and difficulty parameters, Rasch

(1960,1980) explains that this can be conducted in a manner which is consistent with this view of the parameters as representing intrinsic properties of the items and persons: the person parameters and their possible distribution can be eliminated from the estimation of the item difficulties, while the item parameters can be eliminated from the estimation of the person abilities. Furthermore, as Samejima (1977:236) points out, the standard error of estimation associated with an IRT ability estimate is free of any particular person sample; similarly, the standard error of estimation for a given item difficulty is independent of the set of items in which that item happened to appear.

Thus item response models have separable parameters, which, as Gustafsson (1977:15) observes, "... can, at least in principle, be estimated on scales that are independent of the particular sample of examinees studied." Lord (1974a:113) notes that as a result, it is possible to determine the item parameters "once and for all" by pretesting in some convenient group of testees: once a set of items has been calibrated in this way, a person's ability can be estimated from his/her responses to any sub-set of these items, using the maximum likelihood equations (Baker, 1977:171).

In practice, however, Lord (1974a:113) advises against placing too much reliance on the invariance of IRT parameter estimates when there are wide variations between samples, and recommends that the persons used for pretesting should resemble those who will later be tested. It must also be remembered that, as was mentioned in the previous section, estimates are most precise when items and persons are well-matched in difficulty/ability levels: a suitable pretesting sample would therefore be one which contained sufficient persons close in ability to each item difficulty for accurate estimates of the item parameters to be obtained.

Hulin et al. (1983:44) point out that sets of item parameter estimates calculated from different person samples should not in any case be expected to be identical, since sampling fluctuations will normally result in some variation. However, they should in theory exhibit a certain stability, taking into account the error of measurement for each estimate. This applies equally to sets of ability estimates obtained separately for the same testees using different sub-sets of items from a calibrated pool. As Wright (1968:94) explains, even if persons were tested twice using the same test, their scores would be unlikely to be exactly the same each time, because of the measurement error that enters into every testing procedure. He argues, therefore, that the important question is not that of

whether the ability estimates obtained using different tests are identical, but whether they can be said to be statistically equivalent, i.e. whether the differences between them are roughly those that one would expect, given the error of measurement associated with each.

As regards the scale upon which the persons and items are placed using IRT procedures, Subkoviak and Baker (1977:304) and Hambleton et al. (1978:472) note that both the origin and unit of measurement are arbitrarily determined: a reference point for each analysis is customarily provided by setting a mean value either for ability or difficulty on the common scale (Hambleton, 1979:21).

Once the basic scale has been determined in this way, simple linear transformations can be performed to yield scales with the properties required for particular applications (Willmott and Fowles, 1974).

Although the ability scale yielded by IRT estimation procedures is described by Subkoviak and Baker (1977:304) as being " ... a function of the set of items, the group of subjects, and the techniques used to "anchor" or define the origin and determine the unit of measurement", this scale is nevertheless widely acknowledged to possess certain desirable properties. One of these is that it is an equal interval scale (see e.g. Raatz, 1985:61); indeed, it is described by Wright (1968:96) as a ratio scale, since it allows the tester to state precisely how many times more or less able one person is than another.

The implications for testing practice of being able to obtain estimates with the properties outlined here will be considered in Section 2.3.7. It must be stressed, however, that these properties are achieved only if there is satisfactory fit between model and data. Gustafsson (1977:91) emphasises that the basic assumptions of the model must be fulfilled in order for any reasonable estimates of the parameters, or any sensible application of the model, to be possible at all. The concept of model-data fit, and some methods for investigating this, are discussed in Section 2.3.6 below.

2.3.6 Evaluation of Fit

When a mathematical model is found to account adequately for observed instances of the event which it seeks to represent, there is said to be fit between model and data. In the case of an item response model, which is a mathematical model of the relationship between certain characteristics of test-takers and test items, the notion of fit concerns the degree of correspondence between the

model representation of this relationship and the reality as evidenced by a set (or sets) of response data.

The evaluation of model-data fit is an essential part of any application of IRT in that it provides a check on whether the parameter estimates obtained can be treated as plausible. As Traub and Wolfe (1981:413) warn, "... if the model is inherently and grossly wrong for the data, then applications of the analytic results, which most programs will produce regardless of fit, can be nonsensical." Thus the goodness of fit between model and data must be investigated before any use of the statistics obtained from an IRT-based analysis of test data can be contemplated.

As well as providing information about the general quality of the analysis and the usability of the results (Traub & Wolfe, 1981:413), investigations of fit help the tester to identify patterns of unexpected or inconsistent responses at the level of individual persons and items. The importance of this, both in terms of quality of person measurement and with regard to item selection, is considered below.

2.3.6.1 Implications of Person and Item Misfit

As Weiss and Davison (1981:644) explain, "Observed lack of fit for an individual permits the conclusion that the model is an inappropriate means of describing the behavior of that individual on that set of items ...". Where most of the persons in a group have responded largely in accordance with the model's expectations, an instance of person misfit can usually be attributed to anomalous test-taking behaviour of some kind. For example, an unexpectedly large number of incorrect answers by a high ability person may in some cases result from inexperience with the test format, or failure to pay sufficient attention on easy items, while an unexpectedly large number of correct answers by a low ability person may result e.g. from cheating, lucky guessing or unauthorised access to the test paper in advance (Hulin et al., 1983:112-113).

Whatever the underlying cause, a response vector which is inconsistent with an otherwise well-fitting model may indicate that the test, though possibly functioning well for the group as a whole, has failed to provide an appropriate measure of the relevant ability for that particular person (Hulin et al., 1983:111). Indeed, Hulin et al. (1983) discuss person fit under the heading of "appropriateness measurement", a term deriving from the work of Levine and Rubin (1979), in which several indices of person fit ("appropriateness indices") are presented.

Where an aberrant response pattern is identified for a particular item across all the persons in the group (rather than for a person across all the items), this may indicate that the item is flawed in some way, or that it does not tap the same ability as the others in the set, or, if certain systematic inconsistencies in the responses of identifiable subgroups are observed, that the item is biased.

Clearly, such instances of item misfit need to be investigated, since they may have important implications for the quality of the testing procedure, and hence for the validity of any decisions based upon the test scores.

2.3.6.2 Statistical Tests of Fit

A detailed discussion of the various measures of fit which have been developed would be outside the scope of this study, particularly as they are in some cases associated with a specific model, or indeed with a particular parameter estimation procedure. However, an indication is given in this and the two following sections of some approaches to the evaluation of fit, and some suggested methods for checking whether the data satisfy model assumptions.

Traub & Wolfe (1981:414) note that the comparison of theoretical prediction with observed reality is basic to the evaluation of any theory. Accordingly, most tests of goodness of fit in IRT are based on some form of comparison between the expected responses as predicted by the model and the actual responses observed in the data. These tests are in some cases designed for the evaluation of overall fit between model and data, and in others for the detection of misfitting persons and items at an individual level.

In general, checks of observed vs expected responses involve the following preliminary steps: (i) fitting the model to the data, i.e. estimating the person and item parameters in accordance with the item response function specified by the model, and (ii) using these estimates in the item response function to estimate the probability of success for each person (or for each subgroup of a given estimated ability level) on each item in the test. The differences between the observed responses (assigned the value 1 if correct, 0 if incorrect) and these estimated probabilities are referred to as residual differences, or residuals, and these form the basis for a number of statistical tests of goodness of fit.

Gustafsson (1977) notes that statistical tests of goodness of fit, which exist for all the different IRT models, are usually of the chi-square or likelihood ratio type. Chi-square tests based on the analysis of residuals are described in the

work of Wright & Panchapakesan (1969), Mead (1976) and Wright & Stone (1979). The tests of person and item fit implemented in the BICAL program for item calibration (Wright, Mead & Bell, 1980) are of this type; details of these are given in Section 2.4 and in Appendices A.3 and A.4. The Martin-Löf test of fit, described by Gustafsson (1977:51) as a chi-square sum formed from the deviations between observed and predicted frequencies of correct responses within each score group, also belongs to this category.

As Hambleton & Swaminathan (1985:154) explain, likelihood ratio tests involve evaluating the ratio of the maximum value of the likelihood function under the hypothesis of interest (e.g. that the model fits) to the maximum value of the likelihood function under a competing hypothesis. A likelihood ratio test which is frequently referred to is that developed by Andersen (1973), for the evaluation of overall fit to the Rasch model. This is based on the maximum value for the likelihood function calculated using the item parameter estimates obtained from the whole person sample, and the sum of the maximum values of the likelihood function using item parameters estimated separately within different ability subgroups (see e.g. Gustafsson, 1977:48-49).

As their name suggests, likelihood ratio tests are closely associated with maximum likelihood parameter estimation procedures. According to Traub & Wolfe (1981), such tests can provide evidence in support of the decision to reject a model, but are not entirely satisfactory as far as the detection of deviations is concerned. They are, however, considered to be of particular value in studies of the fit of alternative models to the same set of data. Waller (1981), for example, reports on a comparative study of the fit of three different models using, for each item, a likelihood ratio chi-square goodness of fit statistic based on the log likelihood equations for item estimation, and then summing these across items to provide a measure of the overall fit of each model.

Although tests of goodness of fit based on the analysis of residuals may appear to embody an approach different from that of likelihood ratio tests, it can be shown that the two methods are related. As Traub & Wolfe (1981) point out, the likelihood function will be enhanced if the probabilities of success were large for the correct responses made, and small for the incorrect ones. Thus the value of the likelihood function is directly influenced by the magnitude of the residuals: the smaller the residuals, the larger the value of the likelihood function. Traub & Wolfe (1981) further note that a conventional statistical test such as that of Andersen (1973) can be interpreted as indicating whether or not the size of the

residuals would be consistent with random fluctuations within the model.

Likelihood ratio tests can also be shown to be related to tests such as those of Wright and Panchapakesan (1969), involving the discrepancies between observed and expected frequencies of correct answers across ability groups: Mead (1976) views this relationship as arising from the fact that both are based on the principle that all score groups must give statistically equivalent estimates of the item parameters.

This principle appears overtly in work on the evaluation of fit by a number of contributors to the field. Rasch (1960, 1980), for example, suggests carrying out a statistical test to establish whether estimates of item difficulty calculated from different ability groups differ by more than random variation. Wright, Mead and Draba (1976) present a statistic based on the difference between the difficulty estimates for an item calculated in different subgroups of testees; they suggest that their procedure could be of use in detecting item bias, since it in effect indicates whether an item is more difficult in one subgroup than another, taking into account the standard error of each estimate. The F-test of item bias used by Hulin, Drasgow & Komocar (1982) indicates whether the observed ICCs for two subgroups are sufficiently similar for it to be reasonable to treat them as one group, and is thus also based on the requirement that the difficulty estimate associated with each item should remain the same, irrespective of the particular subgroup from whose responses it was calculated.

Thus although Hambleton and Murray (1983:72-73) categorise approaches to the evaluation of fit as though checking for expected model features (e.g. parameter invariance) represented an approach distinct from that of comparing model predictions and observed outcomes, it can be seen that these are in fact closely related.

Methods for assessing person fit in particular are discussed at some length by Hulin et al. (1983). They note that Levine and others have devised several indices of person fit based on the size of the likelihood function: persons for whom no estimated ability parameter yields a relatively large value for the likelihood function will be identified as misfitting. Two other indices of person fit, also devised by Levine et al., belong to what is known as the Gaussian class. As Hulin et al. (1983) explain, one of these is based on a likelihood ratio involving the ability estimate obtained under an item response model of the type discussed here, and the corresponding estimate obtained under the Gaussian model, in which a separate ability estimate is calculated on the basis of the response to

each item. The other makes use of an estimate of the variance of the ability distribution for each person under the Gaussian model: this variance should be close to zero if the person's response pattern is determined largely by a single ability. Where other factors, such as cheating or misunderstanding, have influenced the response vector, the variance will be greater than zero.

It should be noted that a number of authors express misgivings about the use of certain statistical tests of fit. Hambleton and Swaminathan (1985:152), for example, consider it inappropriate to place a great deal of emphasis on statistical tests, particularly those of the chi-square type. They claim (p153) that the chi-square test "has dubious validity" in cases where any of the expected terms has a value of less than 1, and cite the study by van den Wollenberg (1980) to support the contention that the statistic presented by Wright and Panchapakesan (1969) is not, in fact, distributed as a chi-square variable, and that the associated degrees of freedom are not, in reality, as high as they have been assumed to be.

A general problem raised in connection with tests of fit of the chi-square type, however, is that this statistic can be greatly influenced by sample size. Hambleton and Swaminathan (1985:153) note that if the number of observations taken is sufficiently large, the (null) hypothesis that the model fits the data will invariably be rejected. Hulin et al. (1983) further note that if the sample is too small, even gross departures from the model may pass unnoticed.

Indeed, according to Traub & Wolfe (1981:418), inferential measures of goodness of fit in general suffer from this "well-known and insurmountable problem", and it is this which has stimulated work on the development of measures of fit which are descriptive rather than inferential.

2.3.6.3 Graphical Tests of Fit

It is suggested by a number of authors that statistical measures of fit should be supplemented, or even replaced, by graphical methods. As Mead (1976) observes, the earliest tests of fit were of this type: Rasch (1960), for example, suggests plotting pairs of difficulty estimates obtained from different ability subgroups, or, for a more stringent test, plotting the estimates obtained for each score group against the average for the sample as a whole. If the model fits, the points should in both cases form a straight line with unit slope.

Other graphical methods are described by Gustafsson (1977), who suggests plotting, for each item, the observed proportion of correct answers within each

score group against the expected proportion, and by Hambleton (1980), who gives examples of plots of the residuals across ability groups. Graphs of these types are considered to be a useful aid in examining the data for deviations, and in interpreting, and assessing the relative seriousness of, any deviant patterns found. However, while these methods have the advantage of being less sensitive to sample size than statistical methods, they are also considered by Gustafsson (1977) to have a slight disadvantage in that they involve an element of subjective judgement.

Although described by Gustafsson (1977:43) as being among the "more primitive" tests of fit, the groupwise comparison of parameter estimates is nevertheless considered to be a useful indicator of fit. Both Wright (1968) & Lord (1980), for example, compare item parameter estimates obtained from testees at the high and low extremes of the ability ranges in their samples. As Hambleton & Swaminathan (1985) note, such pairs of estimates should, when plotted, show a linear relationship: departures from this would indicate that the model does not hold either for one or both of the subgroups.

Divisions of the testee sample can be carried out using criteria other than ability: an example of the use of this procedure in examining items for possible bias in different racial groups can be found in the work of Hambleton & Murray (1983:80-83). Other criteria for the formation of subsamples, depending on the circumstances and the purpose of the analysis, might be geographic region, high and low performance on some other measure, or course of instruction followed (Hambleton & Swaminathan, 1985:158).

Checks on the test-independence of ability estimates can be carried out in a similar way. A comparison of ability estimates obtained for the same people using the easier and harder halves of the same test is reported by Wright (1968). This involved checking the two sets of estimates for statistical equivalence. Hambleton and Swaminathan (1985), however, recommend plotting the pairs of estimates, noting that the resulting plot should be linear, though with some scatter due to the effects of measurement error. Where a linear relationship is not obtained, this indicates that one or more of the model assumptions is being violated.

Again, the division of items by difficulty is not the only possible approach: as Hambleton and Murray (1983:76) suggest, other meaningful divisions, e.g. according to different content categories within the test, may also prove informative for purposes of checking on the invariance of ability estimates.

The main difficulty arising in the use of the graphical methods mentioned here is that of knowing how widely the points can be scattered about the best-fitting line before the fit between model and data is called into question. Wright & Stone (1979:94-95) deal with this problem by setting up confidence boundaries calculated using the standard errors associated with each pair of estimates.

Another suggested technique for overcoming the difficulty of interpreting differences between estimates presented in graphical form is attributed to Angoff (1982). This involves conducting additional comparisons, using pairs of parameter estimates obtained from subsamples formed by random divisions of the sample of interest. When plotted, these provide an indication of the amount of variation found in the estimates when calculated using similar subgroups: they are therefore of use in deciding at what point the variation in estimates obtained using different subgroups might be considered noteworthy. Thus the function of these additional plots is to provide a point of reference for the interpretation of differences found in the main experimental investigation based on some rational division of data. They are therefore referred to as 'baseline plots' (see e.g. Hambleton & Murray, 1983).

2.3.6.4 Checks of Model Assumptions

Although, in effect, the checks of fit outlined so far provide information as to whether the data satisfy the assumptions required by the model, there are, in addition, various suggested procedures for checking on specific assumptions.

As was indicated in Section 2.3.4.2, the assumption of unidimensionality requires essentially that the items in a test should form a homogeneous set. (Indeed, Gustafsson (1977:73) comments that the question of item fit is concerned not with whether 'items fit the model', but with whether they fit together to form homogeneous scales.) The aim in investigations of the dimensionality of a data set is thus that of detecting possible sources of item heterogeneity.

Factor analysis has frequently been suggested as a suitable means for addressing this question (see e.g. Lumsden, 1976; Hambleton & Swaminathan, 1985). However, according to Gustafsson (1977:10), factor analytic studies are problematic in some respects, the main difficulty being that of selecting a suitable measure of association between the items. A further difficulty identified by Gustafsson concerns the restriction in the applicability of factor analysis when

there is no variation in ability levels within the data set: he notes that item response models can usefully be applied even when all testees have the same ability level. Hambleton (1980), too, expresses reservations about the use of factor analysis for purposes of checking for unidimensionality, noting that it does not necessarily provide a reliable indicator. A number of methods have therefore been proposed for investigating possible sources of item heterogeneity in a more direct way.

One possible threat to unidimensionality for all the response models discussed here is the effect of a time limit, since, as was pointed out in Section 2.3.4.2, this could introduce an additional dimension (speed) into the measures yielded by the test. According to Hambleton and Swaminathan, little attention has been given to this question. However, they set out several methods by which this type of violation of unidimensionality might be detected. One approach would be to ascertain how many testees failed to complete the test, and the number of items they failed to answer. Another, which they attribute to Gulliksen, would be to compare the variance of the number of items omitted with the variance of the number of items answered incorrectly. A third approach, attributed to Cronbach and Warrington (1951), would be to investigate the relationship between scores obtained under a specified time limit and those obtained in unlimited time. Wright and Stone (1979) examine testees' response patterns for accumulations of omitted items towards the end of the test, since these may be indicative of time effects.

A method for detecting the presence of more than one dimension in terms of item content, but not based on any form of correlational analysis, is proposed by Bejar (1980). This involves plotting pairs of item difficulty estimates obtained from the complete data set against those obtained from subsets of the data, each containing items of a particular content area. According to Bejar (p.284), the points should in each case fall close to a line with unit slope if unidimensionality holds: departure from this would indicate that the content area-based subset was tapping a component unique to that content area.

Hambleton and Swaminathan (1985) consider Bejar's method to be among the more promising approaches for checking on unidimensionality. However, Spurling (1987) questions the rationale for this method, arguing that the test-based and item subset-based difficulty estimates are based essentially on the same information (i.e. on the number-correct item scores for the same set of persons), and thus that they cannot be expected to show departures from

unidimensionality. Spurling considers the comparison of ability estimates obtained for the same persons using the complete test and content area-based item subsets to be a more appropriate method for carrying out such a check.

In Gustafsson's (1977) view, the assumptions of unidimensionality and local independence are in any case tested through the models themselves, using tests of goodness of fit. However, the question of whether the local independence assumption is violated by a given data set has also been addressed separately. Both Whitely and Dawis (1976) and Yen (1980), for example, report studies in which items were administered in different contexts (e.g. in different orders), and the effect of this on the item parameters then examined.

The one- and two-parameter models assume that guessing has not affected the probabilities of success. As Hambleton and Swaminathan (1985) point out, it is not possible to determine directly whether or not correct guessing has occurred; they do, however, suggest some methods which might provide information on this matter. These include (a) investigating the performance of low ability testees on the most difficult items, and (b) examining plots of the proportion of correct answers made by each score group on each item in cases where the performance of low score groups is greater than zero. They note that these methods can only be relied upon if the sample contains testees who are low-scoring in relation to the test, and not only in relation to the others in the sample. Failing this, one can do little more than to consider the item format (e.g. the number of distractors in the case of multiple-choice items) and relevant aspects of the administration procedure (e.g. the time limit) to judge whether guessing is likely to have occurred.

An additional assumption made in the one-parameter model is that items do not vary in discrimination. It is therefore sometimes suggested (see e.g. Hambleton & Murray 1983:75) that in applications involving this model, item-test score correlations such as biserials or point biserials should be examined to determine the extent of variation in discrimination.

2.3.6.5 General Remarks

Some authors (e.g. Hambleton & Swaminathan, 1985; Traub & Wolfe, 1981) take the view that checks on assumptions should be carried out in order to see whether a given model is suitable for use with a given data set, the assumption here being that the data set is fixed, and that the chosen model needs to be able to account for it. Wright (1968), on the other hand, views the choice of a model

as being a matter of selecting a coherent approach to measurement, and constructing tests in accordance with its principles. Analysis of fit can then be used in assessing the degree to which the test construction process has yielded a measure with the required properties, and in identifying any inconsistencies in the response patterns of individual persons and items. Lack of fit is seen as providing information about the data rather than as signalling a deficiency in the model.

Gustafsson (1977:62) observes that serious deviations from the model's assumptions will invalidate most attempts to capitalise on the useful features of IRT. It should be noted, however, that, as Hambleton and Cook (1977) point out, the assumptions of a given model will never be satisfied by any data set, and that the most important questions concern the ultimate usefulness and validity of the results. Lord (1968:990) expresses a similar view:

> "The appropriate question is not whether the model holds exactly – this can hardly be expected – but whether it can provide trustworthy approximate answers to important practical questions."

2.3.7 Practical Implications of IRT

As Lord (1980) explains, it is necessary in most practical test development work to be able to predict the statistical and psychometric properties of a test when it is administered to any group, not only to one for whom a suitable standardising sample has previously been tested. As was noted in Section 2.3.1.2, the use of probabilistic models of person-item interaction in IRT makes such predictions possible.

The availability of sample-independent estimates of item difficulty when the data fit the model has a number of important implications for testing practice. One of these, mentioned by Willmott and Fowles (1974), is that the person samples used in item analysis procedures no longer need to be carefully selected as representative of the population for whom the test is eventually intended. Indeed, the main requirement for the sample is simply that it should contain sufficient persons of ability levels close to the item difficulty levels, so that the items can be calibrated accurately. The updating, or re-designing, of tests can also be carried out using person samples differing in ability from the original pretesting samples, without affecting the comparability of the item statistics across groups (Wainer, 1983; Stocking, 1985).

The availability of test-free person ability estimates means that testees can

be administered any subset of items from a calibrated pool, and yet still have their scores expressed on a common scale. As Wright (1968) explains, this makes it possible to measure one person using a hard test, and another person using an easy test drawn from the same pool, and yet compare their estimated abilities on the same scale. Hambleton (1980:78) points out that even if the two persons' scores are identical, their estimated abilities will be different, reflecting the difference in difficulty between the two tests. By the same token, if a group of testees is given two non-parallel tests of the same attribute (drawn from the same pool), each person's ability estimate will be similar for both tests, even though the raw score distributions for the two tests might differ considerably (Hambleton & Cook, 1977).

The contribution which IRT can make to procedures for equating scores from non-parallel tests is clear from this. An example of its use for this purpose can be found in the report by Rentz and Bashaw (1977), on the development of the National Reference Scale for reading in the U.S.A.. They used item analysis and scaling methods based on the Rasch model to produce a single scale which would allow direct comparison of raw scores from any of the 14 published reading tests (each with a parallel form) included in the project, thereby making it possible to use a total of 28 reading tests interchangeably.

Other practical applications of IRT which capitalise on the invariance of parameter estimates are those concerned with test standardisation and norming, item banking and adaptive (or tailored) testing. It is suggested by Willmott and Fowles (1974) that the information needed for setting up norms for a standardised test could be obtained using relatively small samples if IRT methods were used, provided that the persons were fairly well-matched with the items.

The use of IRT for purposes of item banking is one of the most frequently mentioned areas of application (see e.g. Wood & Skurnik, 1969; Choppin, 1976; Wright, 1977a; Pollitt, 1979; Thorndike, 1982a). The function of an item bank is to store a large number of test items with information concerning the content and psychometric characteristics of each, so that the user can select from this a set of items to construct a test which suits his/her requirements. Using IRT item difficulty estimates it is possible to characterise the items in a way that will be stable from one subpopulation of testees to another. As Willmott and Fowles (1974:49) explain, "... tests can then be constructed as desired and two tests containing no items in common yield estimates of attainment which are on the same scale and statistically equivalent."

Thus, as Wright (1977a) observes, the person measures implied by scores on different combinations of items are automatically equated, thereby obviating the need for elaborate equating and parallel form construction procedures. New items for addition to the bank can be calibrated by including them in a test with established items, i.e. by a 'chaining' procedure (see e.g. Pollitt, 1979; Wright & Stone, 1979); thus items can, if necessary, be calibrated in small batches and then placed on a common scale with other items already calibrated. A related procedure, used when tests have not been given to the same samples, is that of test linking (Gustafsson, 1977:101). Using IRT procedures, it is possible to link tests by including in them a subset of common items.

Gustafsson (1977:102) notes that when an item bank of the kind mentioned above is available, "... a large range of measurement problems can be solved with great efficiency and simplicity". One development of the basic item bank idea which has received a great deal of attention in the recent literature is that of adaptive testing. This is a form of individualised testing in which the items administered to each individual are those which are best suited to his/her ability level. Using IRT-based methods, and administering the items by computer, it is possible to re-estimate the person's ability after each item attempted, and then to select from the available item pool the item which is likely to measure most effectively at his/her current estimated level (see e.g. Weiss (1982, 1983) for accounts of possible procedures). In this way, an increasingly precise estimate of ability should be reached as the testee works through the selected items. It is not necessary even for testees to be given the same number of items: testing can stop once an acceptably low standard error for the ability estimate is reached. IRT has thus opened up the possibility of a new form of computer-administered testing procedure which would otherwise not have been feasible.

The parameter invariance feature can also be brought into play in investigations of item bias, as was suggested in Section 2.3.6. It was also indicated that use of an IRT approach allows individual response patterns to be taken into account (in the analysis of person fit), thereby making possible the identification of anomalies which may reflect on the quality of the test procedure. Pollitt and Hutchinson (1987:82) view this as checking the credibility of measurement for each person. As Hambleton and Swaminathan (1985) point out, such an approach acknowledges the fact that individuals are not equally consistent, and makes it possible to monitor the accuracy with which each person is likely to have been measured.

The concepts of item information and test information in IRT also offer important practical advantages, stemming largely from the fact that it is possible to assess the contribution of each item to the test as a whole. Wright and Stone (1979) describe ways in which information curves can be used in constructing tests with optimal properties for the tester's purposes, while Lord (1968) demonstrates the usefulness of test information curves both in test design, and for purposes of comparing different methods of scoring the same test. Hambleton and Cook (1977) mention the advantage of using IRT to compare the relative efficiency of two (or more) tests in measuring the same ability at different points on the ability scale, and Stocking (1985) makes general reference to the value of information functions for describing the measurement effectiveness of tests.

Bejar (1983a) describes IRT as "neutral" as far as the distinction between norm-referenced and criterion-referenced testing is concerned. Hambleton and Cook (1977) cite as one of the major advantages of the IRT ability scale the fact that the ability estimates can be interpreted in terms of probabilities of success on particular items, thereby allowing content-referenced interpretations of scores as well as the more familiar norm-referenced ones. Indeed, a number of authors advocate the use of IRT-based methods in the construction of criterion-referenced tests. Pollitt (1979), for example, suggests the use of IRT-based item banks as a support for criterion-referenced measurement, and Hambleton and de Gruijter (1983) discuss the advantages of using IRT in the selection of items for mastery tests. In the view of Hambleton and Cook (1977), IRT models provide an "excellent underpinning" for the theory and practice of criterion-referenced testing. They suggest that for each person, items could be sampled (possibly at random) from a pool of items pertaining to an instructional objective, and the ability estimates reported on a common scale; criterion-referenced tests could also be constructed so as to discriminate at different levels on the ability continuum.

As Hambleton and Cook (1977) point out, IRT is thus potentially of great use from a practical point of view, in allowing rigorous investigation of testee performance and in providing a framework for the solution of test design problems. There are, however, certain factors which have militated against its widespread use, among these being the cost of the additional computation required, and the relative unfamiliarity of IRT to many of those concerned with constructing and administering tests. There has, in addition, been a great deal of debate concerning certain issues relating to the use of IRT: these form the basis

for the next section.

2.3.8 Issues in the Use of IRT

The issues raised in connection with the use of IRT fall into two main categories: (i) those concerning the desirability of adopting an IRT framework at all, and (ii) those concerning the relative merits of the different models. This section deals with the first of these categories; issues relating to the differences between the models are discussed in Section 2.4.

Hambleton and Murray (1983:71-72) observe that "... it would be incorrect to convey the impression that issues and technology associated with item response theory are fully developed and without controversy". The fundamental question raised is that of whether the assumptions upon which IRT is based are tenable.

2.3.8.1 Assumption of Unidimensionality

The objections concerning the unidimensionality assumption raised by Goldstein (1979, 1980a, 1980b) and Goldstein and Blinkhorn (1977, 1982) are directed in particular at the Rasch model, but apply equally to the other unidimensional response models. Goldstein and Blinkhorn (1977:310), in arguing against the suitability of such a model for the monitoring of educational attainment in schools, claim that the assumption of unidimensionality "... presupposes a highly simplified view of cognitive functioning", and imposes too great a restriction on the selection of items for a test. According to Goldstein (1979:214) this assumption implies, for example, that one could not include e.g. geometrical and algebraic items in the same test if responses to these were believed to be determined by different mental processes. Indeed, Goldstein (1980a:211) considers the unidimensionality assumption to be inappropriate to educational measurement in general, citing as an example the public examination system (in Britain), in which "... some heterogeneous averaging of marks typically is required."

Thus two main strands can be discerned in the debate on the issue of unidimensionality: the first is the question of whether unidimensionality in educational testing is possible, and the second is that of whether it is appropriate.

As far as the first of these questions is concerned, the differences in opinion seem to result from differences in the way in which unidimensionality is conceptualised. As was indicated in Section 2.3.4.2, the requirement for

unidimensionality is sometimes taken to mean that all the items in a test should depend on a single underlying mental process. Under this view, unidimensionality would indeed appear difficult to achieve or to establish.

The interpretation of unidimensionality in terms of the homogeneity of an item set (see e.g. Lord & Novick, 1968:381), is considerably less restrictive, however. Indeed, as was also suggested in Section 2.3.4.2, even a test which appeared to consist of items tapping different abilities may in practice be unidimensional for a given group of persons. An example of such a test, mentioned by Lord (1980:20), is that of an achievement test in chemistry, requiring both skill in arithmetic and knowledge of non-mathematical facts. It is pointed out that if all the testees in the group were of roughly the same level in arithmetical ability (as may be the case e.g. in a college-level class), then the test would function as though measuring a single ability.

Hambleton and Swaminathan (1985:16-17) note that the unidimensionality assumption cannot be met completely, because of the additional cognitive, personality and test-taking factors that enter into test performance. Such factors would include e.g. motivation, anxiety and experience with the test format. However, they consider that the assumption will be met to a sufficient degree provided that performance is determined largely by a dominant ability. According to Hulin et al. (1983:40), the central question as far as practitioners are concerned is therefore whether a test instrument is sufficiently unidimensional to allow the application of IRT.

With regard to the question of whether educational attainment might reasonably be conceptualised in terms of dimensions, there are again differences of opinion. Brown (1980), for example, considers it largely inappropriate to view educational testing in this way, since attainment tests are not concerned with psychological traits. Thorndike (1982:9), on the other hand, would regard such an approach as being likely to be suitable for test tasks that vary widely in difficulty but relatively little in kind, such as tasks designed to measure comprehension of reading passages of increasing complexity. He acknowledges that the notion of a dimension of e.g. 'competence in history' seems somewhat less satisfactory, but suggests that it might nevertheless be possible to conceptualise this as a dimension of performance on which an individual might be placed high or low. Pollitt (1979), though arguing in favour of the use of IRT-based procedures in some areas of educational testing, notes that these are not well suited to tests which confound two or more poorly correlated dimensions, or to tests which are

essentially tests of knowledge. Thus decisions as to whether the notion of unidimensionality is applicable in particular circumstances must, as Hulin et al. (1983:40-41) observe, involve the "careful application of common sense", coupled with knowledge of the trait(s) being measured.

Another aspect of the question of the appropriacy of unidimensionality is raised in Goldstein's (1980a) remarks concerning the requirement for the averaging of scores across tests. As was mentioned in Section 2.3.4.3, many authors would argue against the practice of adding together scores from heterogeneous item sets, or across tests of different abilities, on the grounds that meaningful measurement is possible only if the scores represent measures of 'the same thing'.

The need for homogeneous item sets is not, of course, exclusive to IRT. As Goldstein (1981:185) acknowledges, this has also been one of the central concerns in traditional testing procedures. Wood (1978:30) draws attention to the conflict that has long existed between the need for unidimensionality in testing and the desire to include items measuring a variety of abilities:

> "... in practice achievement test constructors invariably find themselves torn between the conflicting claims of homogeneity and heterogeneity and in the process become thoroughly mixed up. On the one hand they want to 'cover the syllabus' by sampling content according to a specification, while on the other they worry about biserials being above a certain notional figure. The result is an uneasy compromise ..."

Wood, like Goldstein, sees a risk in narrowing the scope of tests in order to fit a unidimensional model, but also points out that to insist on having heterogeneous tests may be to deny ourselves the possibility of coherent measurement.

In some cases, the objections raised in relation to unidimensionality represent a more generalised objection to the restrictions imposed by any theoretical framework in which educational assessment is viewed in terms of measurement. An additional reason for the recent debate on this issue, however, would appear to be that in IRT, the assumption of unidimensionality is stated explicitly; in traditional test theory, on the other hand, it is made, but largely implicitly.

2.3.3.2 Assumption of Local Independence

The explicitly stated assumption of local independence (i.e. the requirement that one response should not influence another) has also attracted criticism (see e.g. Goldstein, 1980b). The objections to this appear to be firstly that violations

of this assumption are difficult to detect, and secondly that it is in itself a restrictive assumption (though again, one which is not exclusive to IRT).

The studies carried out by Whitely and Dawis (1976) and Yen (1980), on the effects of item context on IRT parameter estimates, offer some evidence that item parameters estimated from the same context might be more closely related than those estimated for the same items in different contexts. However, Yen (1980) reports that these differences did not in her study greatly affect the relative sizes of the ability estimates. The potential seriousness of the consequences of violation of the local independence assumption would therefore need to be considered in relation to the type of application envisaged: Yen considers (p.309) that for studies in which the ability scale is not important (e.g. studies involving correlations), context effects on item parameter estimates are likely to be relatively unimportant. Where the ability scale is of greater importance (e.g. in test equating), Yen would advise that the same context be maintained.

2.3.8.3 Stability of Parameter Estimates

Goldstein and Blinkhorn (1977) question the notion of the stability of item difficulties across different testee groups: they would not, for example, expect the difficulties of items measuring mathematical ability to be the same for testees who had been taught 'traditional' maths and those who had been taught 'new' maths. They also call into question the entire notion of item banks, since these make use of difficulty estimates which are assumed to remain constant: they claim that in reality, the difficulties of items may change over time as items increase or decrease in applicability.

It is, however, acknowledged by those who advocate the use of IRT for these purposes, that close monitoring of the behaviour of items over time is necessary. Indeed, Pollitt (1981) considers IRT item statistics to be useful in identifying and monitoring changes in difficulty: an example of such change noted in connection with maths items has been that items involving decimals have become easier in the years since the introduction of decimal currency in Britain.

The question of whether the invariance of parameter estimates which holds in theory is achieved in practice is also raised by Subkoviak and Baker (1977:304), in connection with the ability estimates obtained for the same person using different item sets. They claim that such invariance is "... difficult to obtain in all but highly specialised situations".

2.3.3.4 Model-Data Fit

A number of issues are raised in connection with the fit between item response models and data from educational tests.

Although the choice of the item response function was, at least for the two- and three-parameter models, motivated by empirical evidence available in existing test data, doubt is expressed e.g. by Goldstein and Blinkhorn (1977) that such models, and in particular the Rasch model, can adequately describe real data. Bryce (1981), in responding to this criticism, notes that no supporting empirical evidence is presented by these authors.

Choppin (1976:238) accepts that models "... necessarily portray a simplified and somewhat idealized picture of the real situation", but regards them as being of use in simplifying complex situations. He considers their value to lie "... not only in how well they fit the data, but also in the extent to which they lead to useful results."

The originators of these item response models do not, in any case, claim that these models will always reflect reality: Rasch (1960, 1980), for example, observes that although models need to be applicable, they are not true, and should therefore only be accepted "on trial" rather than definitively. Lord (1980), too, emphasises that a mathematical model should not be expected to hold for every item and every testee, since there are many other factors which can influence the outcome in an unpredictable way: if, for example, a testee becomes tired, ill or uncooperative during a test, or if testees omit items through indifference rather than through inability to answer them, then no model can be strictly appropriate.

Wright (1968) and Willmott and Fowles (1974) take the view that items should be constructed (or selected) in such a way that fit to a chosen model is achieved: they regard this as a way of ensuring that items measure in a common way, and conform to the requirements for 'good' measurement (e.g. unidimensionality, local independence). Items which are found to deviate from these can then be investigated and modified, or discarded. Others, however, fear that items for educational tests may, if this approach becomes generally accepted, be selected on grounds of fit to a model rather than on grounds of content. Goldstein (1979:216) considers that the procedure of discarding ill-fitting items amounts to choosing test content on statistical rather than educational grounds.

An additional criticism made by Goldstein and Blinkhorn (1982) in relation to fit is that it is possible that an apparently good fit might disguise evidence of there being more than one dimension involved: they warn that if testers restrict themselves to a one-dimensional model when the reality is multidimensional, it will not be possible to make sound inferences about the data.

There is also disagreement with regard to the measures of fit which should be used: in some cases this concerns the validity of particular methods (as e.g. in the work of van den Wollenberg (1980), mentioned in Section 2.3.6.2), and in other cases their stringency (see e.g. Goldstein & Blinkhorn, 1982).

Goldstein (1980a) observes that even if satisfactory fit to a model were established for a given set of items, it would not necessarily follow that they were measuring anything meaningful. He cites an article by Wood (1978), in which a set of random coin-tossing data was shown to fit the Rasch model, in support of this argument.

It is indeed the case, as Stenner, Smith and Burdick (1983) emphasise, that a response model does not embody a construct theory. They note (p.308): "Nothing in the fit between response model and observation contributes to an understanding of what the regularity means."

The dismissive comments of Goldstein in this regard give the impression that response models have been suggested for use in seeking meaning in random collections of items. Willmott and Fowles (1974), on the other hand, would take the view that given a set of items which, to informed judges, appear to constitute a sensible and coherent test of the content area of interest, the fit of the resulting data to a chosen response model, while not (necessarily) shedding any light on the construct itself, can nevertheless provide an indication of how consistently the items function as a measuring instrument. It can also point to irregularities for which explanations might be found. It is not, therefore, claimed that IRT-based analysis establishes test validity, but rather that, like traditional item analysis, it provides information concerning quality of measurement.

2.3.8.5 Misuse of IRT-Based Procedures

Traub and Wolfe (1981:377) see a certain danger in the uninformed use of computer programs for IRT analysis in educational testing, and comment that the dramatic increase in the number of applications of IRT to educational testing problems "... is undoubtedly more a consequence of computer program

availability than of understanding fostered by the available expositions ...". A similar point is made by Hambleton (1980:93), who observes that the widely-documented potential of IRT for solving a variety of measurement problems is not guaranteed simply by processing test results through a computer program.

Another fear which has been expressed is that use of IRT procedures may lend spurious precision to educational tests (see e.g. Tall (1981) who, like Goldstein, directs his criticisms in particular against the Rasch model). Reservations such as these concern the possible misuse of methods of analysis deriving from IRT rather than any shortcomings of the theory itself. Hambleton and Murray (1983:91-92), in an attempt to counter any misconceptions, summarise the appropriate use of IRT as follows:

> "Item response theory is not a magic wand to wave over a data set to fix all of the inaccuracies and inadequacies in a test and/or the testing procedures. But, when a bank of content valid and technically sound test items is available, and goodness of fit studies reveal high agreement between the chosen item response model and the test data, item response models may be useful in test development, detection of biased items, score reporting, equating test forms and levels, item banking, and other applications as well."

2.4 The Rasch Model

In the first part of this section, the relationship between the Rasch model and the two- and three-parameter models is considered. Attention then turns to a more detailed description of the Rasch model, and of some of the analytic procedures deriving from it.

2.4.1 Relationship with the 2- and 3-Parameter Models

Athough the Rasch model is sometimes described as a special case, or a restricted form, of the two-parameter model, it should be noted that it was not conceived of as such by Rasch. As Douglas (1982:132) explains:

> "Rasch's model was never the consequence of simplifications to a higher-order model but the necessary result of fundamental measurement principles, principles of such generalizability that they could be applied to measurement situations well beyond those rather narrow ones conceived of by many psychometricians working on the other side of the Atlantic; ..."

The measurement principle with which Rasch was primarily concerned was that of 'specific objectivity' in comparisons, i.e. comparisons of persons which did not

depend on the particular items used, and comparisons of items which did not depend on the particular persons tested. Douglas (1982:132) points out that Birnbaum (1968) appears to have chosen the logistic model for reasons of mathematical convenience, to avoid the estimation problems associated with the normal ogive model; for Rasch, on the other hand, the choice of the logistic model was a mathematical necessity if the property of specific objectivity was to be achieved.

Much discussion of the three logistic models introduced in Section 2.3.3.2 has centred on the implications of the difference in the number of item parameters, and two main viewpoints emerge: (i) that the Rasch model, in specifying only one parameter, is too restrictive to describe real data, and that the more complex models are therefore to be preferred, and (ii) that the Rasch model, by virtue of its simplicity and elegance, offers a number of important conceptual and practical advantages over the more complex models.

Hambleton and Swaminathan (1985) note that it has been suggested that the three-parameter model should be generally adopted, since, as the most complex and hence the most general of the unidimensional models in common use, it should in theory result in better fit to test data than the one- and two-parameter models. The inclusion of the 'guessing' parameter is thought by some to be particularly appropriate to applications involving e.g. true-false and multiple-choice items.

The Rasch model is thus sometimes regarded as a less sophisticated version of the two- and three-parameter models. Such a view is implied by Weiss (1983), when he states that the Rasch model enjoyed a certain popularity in the 1960s because some of its procedures for estimating item and examinee parameters could be implemented without the aid of a computer, but that when procedures for estimating the second and third item parameters became available, the "more realistic" two- and three-parameter versions came to be used in preference.

The absence of a discrimination parameter in the Rasch model implies that all items discriminate equally (see e.g. Bock & Wood, 1971). Thorndike (1982a:11) points out that the usual process of item selection tends to eliminate items with low item-trait correlations, i.e. those with particularly flat ICCs, and therefore tentatively suggests: "Perhaps we do not strain reality too much if we assume that the slopes are all equal." He accepts, however, that this assumption frequently represents an oversimplification of reality. Traub and Wolfe (1981),

too, note that it is an empirical fact that items in some tests differ in discrimination. Birnbaum (1968:402) presents evidence of this, in the form of the means and standard deviations of item-test score biserial correlations for a sample of over 3,800 testees on several different multiple-choice tests.

Since the more complex models take into account a larger number of factors which may have influenced the data, it is to be expected that they will show better fit in many cases. However, a completely different view of the desirability of incorporating these additional item parameters is presented by Wright (1968:100), who regards the factors which they represent as having no place in measurement. He writes:

> "We can construct tests in which guessing plays a big part, in which items vary widely in their discrimination, and in which the answer to one item prepares for the next. But do we want to? Not if we aspire to objective mental measurements. If we value objectivity, we must employ our test-constructing ingenuity in the opposite direction."

Brink (1971:101) points out that the inclusion of items with different discriminations in the same test results in a measuring instrument with varying units of measure, and Pollitt (1979:59) draws attention to a further problem which arises when response models allow for varying discriminating power: in cases where ICCs cross, the relative difficulty of the items will not be the same for persons of all ability levels. He notes that "It is to avoid this conceptual and practical difficulty that the restriction of equal discrimination is imposed." A related point, mentioned by Wright (1977b), is that the discrimination values estimated for a set of items will reflect the distribution of ability in the testee group, i.e. unlike the difficulty estimates, they will be sample-dependent. Indeed, Wright and Stone (1979:ix) state that "... *only* item difficulty can actually be estimated consistently from the right/wrong item response data available for item analysis."

With regard to the practical consequences of the different models, Thorndike (1982a) notes that since the three-parameter model requires estimation of so many parameters, very large data samples are needed if the estimates are to show a satisfactory degree of stability across groups, and the estimation procedures require the availability of high speed and high capacity computing facilities. According to Lord's recommendations, pretesting of items would need to be based on samples of over 1,000 testees, a requirement which Thorndike considers to be somewhat unrealistic in many cases.

Hulin et al. (1983:60) explain that in estimating person ability according to the three-parameter model, the optimal weighting of each item will vary as a function of ability, so that the same response to an item will receive different credit for individuals of different abilities: it is in this way that the likelihood of correct guessing is reflected in the person's score. In the two-parameter model, a sufficient statistic for estimating the person parameter can, according to Traub and Wolfe (1981:390), be obtained by summing the discrimination parameters for only those items which the person answered correctly. This, as they point out, assumes that the discrimination parameters are either known, or that they can be accurately estimated, both of which require the use of large data sets. Where discrimination parameters are unknown, and the sample contains fewer than 200 testees, they note that Lord advises use of the Rasch model instead.

Thorndike (1982a) notes that the computational simplicity of the Rasch model is such that a first approximation to the ability and difficulty estimates can be calculated by hand, and that since there is only one item parameter to be estimated, stable estimates can be achieved using smaller samples than those required by the more complex models. Furthermore, it is only in the case of the Rasch model that the number-correct person and item scores are sufficient statistics for the estimation of person ability and item difficulty respectively. Thus it is only in this case that the estimation of ability requires no analysis of the response pattern, but is based simply on raw score. As Waller (1981) points out, this has the important practical advantage that future users of a test calibrated using Rasch-based procedures need not calculate ability estimates by computer: they can simply read off the appropriate ability estimate from a raw score-to-ability conversion table. This is not so for the two- and three-parameter models, which require that correct responses be weighted differently for each person, with the result that the same raw score would correspond to various different ability estimates, depending on which particular items the person had answered correctly.

Lord (1983) observes that in practice the number-correct score (as used in the Rasch model) can provide up to 95% of the information given by the weighted sum (as used in the two-parameter model), and that if the discrimination estimates are sufficiently inaccurate, the number-correct score will in fact be more informative.

Lord and Novick (1968) consider that when its basic assumptions are satisfied, the Rasch model offers a mode of analysis of great simplicity and power. Lord

(1980:190) warns that if these assumptions are not met, "... then use of the Rasch model does not provide estimators with optimal properties"; he points out, though, that when the person sample is small, the Rasch parameter estimates may be more accurate than those obtained using the three-parameter model, even when the latter model holds and the Rasch model does not.

Willmott and Fowles (1974) acknowledge that the assumptions of equal discrimination and minimal guessing may seem stringent, but note that in practice some departure from these can be tolerated. Pollitt (1979) observes that the Rasch model has been shown to fit various kinds of real test data, even though the tests were not constructed in accordance with its requirements. He further notes that in practice, most items which would be considered acceptable using traditional criteria would also be accepted by the Rasch model.

Notwithstanding the claims that more complex models are required in order to describe real data, Choppin (1976) is convinced, on the basis of experience, that for tests of "typical homogeneity", the Rasch model fits well enough to be useful. Indeed, for purposes of item banking, he regards the Rasch model as being the most useful tool currently available for dealing with the complexities involved.

Such views are at variance with those of Whitely (1977:229), who claims that studies in which a "reasonably stringent" test of fit to the Rasch model is applied "... are notable for the frequency with which the model is found to be inappropriate ...". She also considers it possible that the selection of items with uniform ICC slopes to conform to the requirement of equal discrimination might alter what is actually measured if unidimensionality does not strictly hold (p.233). A different perspective on this matter is provided by Gustafsson (1977:18), who considers the Rasch model to be "safer" than the two- and three-parameter models in its potential for detecting important deviations from the main assumptions. He raises a point originally put forward by Mead, concerning the risk, when using the less restrictive models, of inadvertently 'explaining away' threats to the unidimensionality assumption by treating them as instances of varying item discrimination. Thus it would appear that the more complex models, though seeming to account better for real data, may in fact mask important violations that a simpler model might expose.

In view of the desirable features of the Rasch model, a number of authors advocate its use in preference to the more expensive and computationally more cumbersome procedures based on the two- and three-parameter models.

Perhaps one of the strongest arguments advanced in its favour, though, is that the assumptions made by the Rasch model are in fact those upon which most tests are already based. As Pollitt (1979:59-60) explains:

> "Whenever a test is used to provide a single score for each person, unidimensionality is implicitly assumed; and whenever this score is simply the number of items correct, equal discrimination and hence the Rasch model are similarly assumed."

The two- and three-parameter models, in giving greater weight to the more discriminating items, thus represent a quite different approach from that traditionally adopted in testing. The Rasch model, on the other hand, shares its basic assumptions with the traditional approach, but makes these assumptions explicitly rather than implicitly.

2.4.2 Development and Formulation

The Rasch model with which this study is concerned is in fact one of three models developed by the Danish mathematician and statistician Georg Rasch, and presented in Rasch (1960, 1980).

Two of these models were developed as tools for use in assessing different aspects of reading ability: one was concerned with the number of misreadings made in an oral reading test, and the other with oral reading speed. The third model, which is the one now widely referred to as the Rasch model, is the one-parameter logistic model for dichotomously-scored items, discussed in earlier sections. Lord and Novick (1968) note that the models proposed by Rasch have certain common characteristics: each has two parameters, one identified with person ability and the other with item (or test) difficulty, and the ability and difficulty parameters are in each case 'separable' in that they can be estimated independently, in a manner which Lord and Novick describe as analogous to the estimation of parameters in a two-way factorial analysis of variance with no interaction terms.

Details of the development of each of Rasch's models are given by Wright (1980). The model of interest here was developed in the early 1950s, in the course of work on the other models, and tried out on an intelligence test with which Rasch had previously been involved. This test was found not to conform to the model, largely, it appeared, because it consisted of groups of items involving different types of content. A new test, consisting of subtests each measuring a particular ability, on the other hand, showed good fit, and it was this

which increased Rasch's confidence in the applicability of the model. Wright (1980) notes that it was not, however, until 1960 that Rasch's work on item analysis became known outside Denmark.

In developing these new models for test data, Rasch's aim was, as was suggested in the previous section, to put into operation concepts of measurement which were radically different from those used in the traditional theory (Rasch, 1980:4). In particular, he wished to eliminate the role of populations in assessing the levels of individuals. It is, as Rasch explains, through the definition of the two types of parameter (one for persons and one for items) that these new concepts of measurement were introduced. He states the requirements for these individual-centred statistical techniques, as opposed to group-centred ones, as follows:

> "Individual-centred statistical techniques require models in which each individual is characterized separately and from which, given adequate data, the individual parameters can be estimated. It is further essential that comparisons between individuals become independent of which particular instruments - test items or other stimuli - within the class considered have been used. Symmetrically, it ought to be possible to compare stimuli belonging to the same class - "measuring the same thing" - independent of which particular individuals within a class considered were instrumental for the comparison." (Rasch, 1980:xx.)

The above is Rasch's definition of the property of specific objectivity, to which reference was made in the previous section. Wright (1980:ix) notes that this had been set down by Thurstone in the 1920s as one of the requirements for valid measurement, and explains that objective measurement in this sense requires not only measuring instruments which can function independently of the objects measured, but also a response model in which the instrument and object effects can be separated. The formulation of models with separable person and item parameters represents a major development in psychometric theory, and Rasch's contribution in this field is viewed e.g. by Loevinger (1965) as having been outstanding.

Since people's performance on test items is not always consistent, Rasch considered it appropriate, in formulating his model, to express how easily an item is answered correctly by ascribing to each person a probability of success on each item (Rasch, 1980:73). This probability is determined exclusively by the person's ability level and the item's difficulty level: it is assumed that no other factors exert any influence on the chances of success. Wright and Stone (1979:10-11) note that it is possible to think of various disturbing influences

which might interfere with the expression, and hence observation, of ability; one could also think of additional item characteristics, such as discrimination and vulnerability to guessing, which might affect people's responses to items. They argue, however, that if one wishes to measure a person's ability, then it is reasonable to view his/her test performance as being dominated by that ability. Similarly, they consider it reasonable to regard the difficulty of an item as being its dominant feature as far as people's responses are concerned. They note that the assumption that person responses are dominated by item difficulties and person abilities is already commonly made, since it is implicit in the use of unweighted scores as test results (Wright & Stone, 1979:11).

The probabilistic relationship specified by the Rasch model is such that a high ability person is viewed as having a greater chance of success on any item than a lower ability person, and such that any person has a greater chance of success on an easy item than on a more difficult item. In Wright's formulation, the probability of success is governed by the difference between the person's ability and the item's difficulty (see e.g. Wright, 1968; Wright, 1977a; Wright & Stone, 1979).

If person ability is greater than item difficulty, the probability of success is given as greater than .5. If, on the other hand, item difficulty is greater than person ability (i.e. if the item requires more of the relevant ability than the person possesses), then probability of success is given as less than .5. The case in which person ability and item difficulty are equal is assigned a probability of .5, so that the person is viewed as having equal chances of success and failure on the item.

The choice of a function to represent the probabilistic relationship set out above needs to take account of the fact that the probability must be in the range 0 to 1. The difference between ability and difficulty, which can take any value between minus infinity and plus infinity, can be brought into the range 0 to plus infinity by expressing it as a power of the natural constant e. Using the same notation as in Section 2.3.3.3, this can be written $\exp(b_v - d_i)$, where b_v is the ability of person v, and d_i is the difficulty of item i. In order to bring this exponential expression into the range 0 to 1, it can be expressed in the form of a simple logistic function, as follows:

$$\frac{\exp(b_v - d_i)}{1 + \exp(b_v - d_i)}$$

Thus the probability of success for person v on item i, given the ability of person v, b_v, and the difficulty of item i, d_i, can be written:

$$p(x_{vi} = 1 | b_v, d_i) = \frac{\exp(b_v - d_i)}{1 + \exp(b_v - d_i)}$$

where x_{vi} is the response of person v on item i, and $x_{vi} = 1$ indicates a correct response. The form of the Rasch model shown above is that given by Wright and Stone (1979:15), though again, Greek symbols are not used here.

The Rasch model is also sometimes presented in a log-odds form.

$$\text{odds of success} = \frac{\text{probability of success}}{\text{probability of failure}} \quad \text{i.e.} \quad \frac{\text{probability of success}}{1 - \text{probability of success}}$$

Substituting the probability of success as defined by the Rasch model shown above, we obtain

$$\frac{\dfrac{\exp(b - d)}{1 + \exp(b - d)}}{1 - \dfrac{\exp(b - d)}{1 + \exp(b - d)}}$$

which reduces to exp(b - d). Thus in the Rasch model the odds of success for a person on an item are defined as exp(b - d). By an analogous procedure, it can be shown that a person's odds of failure are exp(d - b). Taking the natural log of exp(b - d) and exp(d - b) gives (b - d) and (d - b) respectively. Thus the natural log odds of success are given by (b - d), and the natural log odds of failure by (d - b).

The units in which these log odds of success are expressed are known as 'logits'. A person's ability in logits is his/her natural log odds of success on an item taken as having a difficulty value of zero, and an item's difficulty in logits is the natural log odds of failure for a person with an ability value of zero (Wright & Stone, 1979:17). The parameter estimation procedures outlined in Section 2.4.3 yield person ability and item difficulty estimates in logits.

The way in which the Rasch model allows comparisons of persons to be made independently of the item used to compare them is illustrated by Wright, Mead and Bell (1980:3). As has already been shown, log odds of success are defined in the Rasch model as (b - d). Thus the log odds of success for person v on item i can be written

$$b_v - d_i = \ln \left| \frac{p_{vi}}{(1 - p_{vi})} \right|$$

where p_{vi} = the probability of success for person v on item i. For a different person, u, the log odds of success on item i are

$$b_u - d_i = \ln \left| \frac{p_{ui}}{(1 - p_{ui})} \right|$$

The following comparison can be made of persons v and u:

$$b_v - b_u = \ln \left| \frac{p_{vi}}{(1 - p_{vi})} \right| - \ln \left| \frac{p_{ui}}{(1 - p_{ui})} \right|$$

It can be seen from this that the two persons are being compared without reference to the difficulty parameter, d_i, and that the difference between them will therefore be the same no matter which item is chosen for the comparison.

Similarly, two items can be compared without reference to the person parameter, and again the difference between them will be the same, irrespective of the particular person used in the comparison.

2.4.3 Analytic Procedures

2.4.3.1 Estimation of Ability and Difficulty Parameters

Wright and Stone (1979) describe three methods for parameter estimation under the Rasch model: PROX, which can be done by hand, UFORM, which is done by hand with the aid of tables, and UCON, which requires the use of a computer. Although these methods vary in complexity and in accuracy, they all yield estimates of the person abilities and item difficulties with their modelled standard errors.

Brief outlines of two of these procedures will be helpful here: PROX, on the grounds that "... it illustrates most of the principles underlying Rasch calibration and measurement" (Wright & Masters, 1982:61), and UCON, which is the estimation procedure used in the data analyses presented later in this study.

It should be noted that before any estimation procedure can begin, it is necessary to remove from the data set any persons with perfect or zero scores, and any items to which responses are either all correct or all incorrect. It is not possible to place such persons and items on the ability/difficulty scale, since all that is known in such cases is that the testees were either too able or not able enough for the set of items administered to them, and that the items were either too easy or too difficult for the group of persons who took them.

originated by Cohen (1979),

PROX (or 'normal approximation estimation'), is designed to approximate the results of more exact, but more complex, estimation procedures (Wright & Masters, 1982:60). As Wright and Stone (1979:21) explain, it requires the simplifying assumption that the person abilities and item difficulties are (approximately) normally distributed.

The first step in the PROX item difficulty estimation procedure is to group together the items according to the number of correct responses made to each. For each item score group, the proportions of correct and incorrect responses are calculated. Initial difficulty estimates are obtained by dividing the proportion incorrect by the proportion correct, and then taking the natural log, to give difficulties in the form of logit incorrect values. Wright and Stone (1979:25, 27) note that these estimates, unlike the traditional facility values, are on an equal interval scale. In order to centre this scale, the mean and variance are then calculated for the initial estimates obtained, and the mean value is subtracted from each estimate, thereby setting the mean item difficulty to zero. Since these values will still be dependent on the ability of the calibrating sample, it is necessary to correct them for the spread of abilities within this sample. First, however, initial person measures must be obtained.

The initial ability estimates are given by grouping together persons with the same raw score, calculating the proportions of correct and incorrect answers for each score group, dividing the proportion correct by the proportion incorrect in each case, and then, as before, taking the natural log. The mean and variance for this set of logit correct values then need to be calculated, for use in the final adjustments which are made to both sets of estimates.

The item difficulty estimates are freed from the effects of the spread of abilities in the person sample by multiplying each one by an expansion factor which takes account of the ability dispersion of persons. The calculation of this expansion factor is based on the variances of the initial ability and difficulty estimates, and on the scaling factor 1.7, which, as Wright and Stone (1979:21)

explain, brings the logistic ogive into approximate coincidence with the normal ogive. The ability estimates are also adjusted to correct them for test 'width', i.e. the spread of difficulties in the test. The expansion factor by which each of the initial ability estimates is multiplied is analogous to that used in adjusting the item difficulties, but this time accounts for the difficulty dispersion of items.

It should be noted that the difficulty and ability estimates corresponding to any possible number-correct item or person score can be calculated: it is not necessary for a given score actually to have been observed in the data in order for the transformation to logit difficulties and abilities to be performed. Clearly, in practice, it will be the raw score-to-ability conversion which will be of interest, since future testees may gain scores not observed in the original sample.

The standard error associated with each PROX difficulty and ability estimate can be calculated in the manner shown by Wright and Stone (1979:44), As was mentioned in Section 2.3.5.2, the standard errors provide an indication of the amount of information available for use in making the item and person measures. As was also mentioned, in the Rasch model the information available in any response is given by the person's probability of success on the item multiplied by his/her probability of failure. Thus information will be at its maximum when probability of success is equal to probability of failure, i.e. when both are .5. This, it will be remembered, is the probability assigned to the case where the person's ability is equal to the item's difficulty. Thus information is greatest, and the standard error smallest, when the person and item are matched in ability/difficulty.

The UCON parameter estimation procedure (or 'unconditional maximum likelihood estimation') was initially developed for the Rasch model by Wright and Stone (1979:62-65); an adapted version of the algorithm given by Wright and Stone may be found in Appendix A.1.

In summary, the UCON procedure involves defining initial estimates (based on the number-correct scores) of each ability and difficulty, and then using these as the starting point for an iterative procedure to solve the equations necessary to maximise the likelihood of obtaining the observed response matrix. As Wright and Masters (1982:61) note, the UCON procedure, in estimating the ability and difficulty parameters simultaneously, does not take full advantage of the separability of parameters allowed by the Rasch model; the reason for this is that the equations which need to be solved are implicit with respect to ability and

difficulty (Wright et al., 1980:6), i.e. the ability and difficulty parameters are inextricably combined within them. Wright and Stone (1979:65) note that the presence of the ability parameters in the likelihood equation has been shown to result in biased estimates of the item difficulties. They add, however, that this can be compensated for by multiplying each difficulty estimate by the scaling factor [(L - 1)/L], where L = the number of items. This correction is applied in the UCON procedure implemented in the BICAL computer program (Wright et al., 1980). In this program, the ability estimates are also corrected for bias, using the scaling factor [(L - 2)/(L - 1)]. The standard errors for the UCON parameter estimates are calculated as shown in Appendix A.2.

2.4.3.2 Measures of Fit

Some methods for evaluating the fit of response models to test data were mentioned, in general terms, in Section 2.3.6. Since reference will be made in later chapters to the statistical tests of fit included in the BICAL program (Wright et al., 1980), these are summarised below.

The information-weighted total fit t-statistic calculated in BICAL for each person and each item is based on the comparison of the actual outcome of each person-item encounter with its expected value according to the model. The residuals are then squared and summed for each person across all the items, and for each item across all the persons; each sum is divided by its model expectation, (Σ p(1 - p)), to form a weighted mean square statistic. This is then converted to a t-statistic which should in theory have a mean of zero and a standard deviation of 1.

The BICAL between-group fit t-statistic involves calculating the squared standardized residuals for the responses made on each item within different ability subgroups, and converting these to a mean square across subgroups for each item. The mean square is then expressed as a t-statistic which, according to Wright et al. (1980:11), tests whether the observed ICCs correspond with the expected ICCs, i.e. whether they have a common shape and slope.

Details of the methods of calculation for these fit statistics are set out in Appendices A.3 and A.4; examples of their application to data will be discussed in Chapters 4 and 5.

CHAPTER 3
USE OF RASCH ANALYSIS IN FOREIGN LANGUAGE TESTING

In this chapter, published studies involving the application of Rasch analysis in the area of second/foreign language testing are first surveyed. Background issues relating to the applications reported in this study are then discussed.

3.1 Reported Applications

Attention is restricted here to studies concerned with second/foreign language tests consisting of dichotomously-scored items, and using Rasch-based methods of analysis; applications relating e.g. to partial credit and rating scale analysis, to other response models, or to testing in the mother tongue, are therefore not included.

The studies described here are categorised according to whether they are concerned primarily with:

1. Investigations of the usefulness of Rasch analysis compared with traditional test analysis;
2. The use of Rasch-based methods in the development of particular measures of, or systems of measurement for, achievement or proficiency in a second/foreign language;
3. The use of Rasch analysis in investigating the nature of sets of language test data.

3.1.1 Comparisons of Traditional and Rasch Methods of Analysis

Comparisons of the results of traditional and Rasch analyses are presented by Henning (1984), Perkins and Miller (1984) and Cziko and Lin (1984).

Henning's (1984) comparison of traditional and Rasch-based item selection procedures was based on the responses of 108 adult learners of English on a 48-item multiple-choice reading comprehension test. The traditional statistics computed for each item were the facility value, item variance, and point biserial correlation. Items were discarded if (a) the facility value was .9 or above, (b) the variance was .1 or below, or (c) the point biserial was .2 or below. Classical reliability indices (K-R20 and K-R21) were computed both before and after discarding the 8 items considered deficient according to these criteria. The modified version of the test was deemed satisfactory from the point of view of reliability.

The Rasch item statistics used in Henning's study were the difficulty estimate and standard error, and the total fit mean square index of fit. Of the 8 items judged least satisfactory in terms of fit, only 4 had also been considered inadequate in the traditional analysis. The selection of items on the basis of fit to the Rasch model was found to bring about a greater increase in the K-R20 indices than selection according to traditional criteria. In view of this, and taking into consideration also the advantages of the Rasch ability/difficulty scale, and the usefulness of the additional information yielded by Rasch analysis (e.g. person fit statistics, standard error for each ability and difficulty estimate), Henning concludes (p.132) that the Rasch approach offers the test developer "numerous advantages".

The stated purpose of the study by Perkins and Miller (1984) was to compare the number of "weak" items detected by traditional and Rasch analysis, and to use the Rasch results in defining an ESL reading variable, and in checking on the suitability of this definition. Perkins and Miller's analyses were performed on the responses of 88 adult learners of English on Henning's 48-item multiple-choice reading comprehension test. The traditional indices computed were item facility values, discrimination indices (using the highest- and lowest-scoring 28% of the sample), and point biserial correlation coefficients. Criteria for the rejection of items were (a) a facility value of less than .33 or greater than .67, (b) a discrimination index of less than .67, and (c) a point biserial of less than .25. A further index, referred to by the authors as an "internal construct validation statistic" was computed, using the same subjects' scores on a 50-item multiple-choice grammar test. This involved calculating the point biserial correlation between each reading item and the total scores for the grammar test, with the expectation that if the reading items possessed construct validity, this second set of point biserials would be lower than those computed using the total score on the reading test (assuming, of course, that the grammar items were not in reality measures of reading).

Perkins and Miller report that the Rasch item fit statistics identified a larger number of "weak" items than any of the traditional indices; they do not explain, however, why they chose to reject not only the 4 least well-fitting items (i.e. those with the highest total fit t-statistics), but also the 9 items with the lowest total fit t-values. Although, as Wright et al. (1980:85) explain, cases of extreme fit may need to be investigated, it seems inconsistent to reject items on grounds of extreme discrimination as indicated by the fit statistics when no upper limit was set for the traditional indices of discrimination. Of the 13 items identified by the

fit statistics, 6 were also identified by at least one of the traditional indices; the differences in the remaining items are not investigated, however, and 'reasons' for preferring the item fit statistic to the traditional difficulty and discrimination statistics are given in terms of the advantages of the Rasch difficulty scale and the disadvantages of the point biserial, rather than with reference to the fit statistic or to the particular items identified by it. Furthermore, it is implied in the conclusion to this study that Rasch analysis is to be preferred for having identified more items than the traditional indices. It would be unusual, however, to carry out item selection on the basis of any one of these indices in isolation: if the facility values and point biserials had been considered together, 17 items would have been rejected, i.e. 4 more than using the fit statistics.

Both Henning (1984:131) and Perkins and Miller (1984:30) take advantage of the possibility offered by the Rasch approach of displaying the item difficulties matched against the person abilities in their samples: this is used in both studies to identify points in the ability range at which additional items are needed in order for persons at those levels to be tested efficiently.

Henning (1984), having used the person fit statistics and standard errors of ability estimates to identify persons whose scores may not be valid, suggests possible reasons for the particular response patterns observed in these cases. Among the influences thought to be operating are guessing, particularly at low ability levels, and test-taking anxiety.

Perkins and Miller (1984) make no mention of person fit, but examine the content of items identified as being the easiest and hardest items in the test, with the aim of determining "... how well the 48 items succeed in defining a variable and exactly what that variable seems to be" (p.29). Although differences are noted between items at the two extremes (the easy items appearing to require processing of explicitly stated information, while the hard ones involve paraphrases and inferences), the claim that the Rasch model allowed the authors to determine what the measured variable seemed to be is misleading: the same items would have been identified as easiest and hardest by their facility values, since Rasch item difficulties maintain the order of the observed proportion-correct item scores, and, as was pointed out in Chapter 2, analysis according to an item response model does not, alone, imply construct validation.

In the study by Cziko and Lin (1984), three different approaches, including classical item analysis and reliability estimation, and Rasch analysis, were used in the investigation of proficiency scales resulting from (a) a modified dictation test

(in which segments of varying lengths were treated as single items) and (b) a "copytest" version of the same passage (in which the segments were presented visually, for restricted lengths of time, instead of orally). These tests were administered to 67 adult learners of 4 proficiency levels, and to a small number of native-speaking undergraduates (17). The results of the analyses appeared to support the findings of an earlier study by Cziko, in which it was concluded that these tests offered convenient procedures for obtaining reliable and valid proficiency scales. It was noted also that in all but one of 20 comparisons carried out, the Rasch log ability scores showed higher correlations with other measures of proficiency (the Test of English as a Foreign Language and the Illinois English Placement Test) than the raw scores.

The conclusions drawn by Cziko and Lin are that the dictation and copytest scales are amenable to Rasch analysis, since guessing is not involved, and since the 'items' were mostly found to have consistently high discriminating power. They further note that although in general the same 'items' were identified as suspect, regardless of the approach used, the point biserial correlations were influenced by item difficulty - giving low coefficients to very easy or very hard items - while the Rasch indices of fit were not affected in this way. They therefore consider the Rasch indices to be preferable.

3.1.2 Use of Rasch-Based Methods in the Development of Language Tests and Measurement Systems

Henning (1986) outlines some of the problems involved in producing proficiency or placement tests for use in medium- to large-scale language teaching operations. He notes that where there is a single, re-usable test, problems which are likely to arise are those of test security and practice effects. The production of alternate forms of a test is also problematic: simply administering the same content in a different order does not seem satisfactory, but the establishing of equated forms by traditional means is a costly and time-consuming process, requiring that two similar tests be administered to the same large sample.

Henning recommends instead the development of Rasch-based item banks for use in language proficiency and placement testing. He reports on such an application at U.C.L.A., where the English as a Second Language Placement Examination forms the basis for an item banking project, using the database management software dBASE II for the storage and retrieval of the calibrated items. A reason given for the choice of the Rasch model in this project is that item banking can be carried out using fewer examinees than for other models.

The U.C.L.A. item banking project is not yet fully developed; however, Henning notes that once such a database is in operation, items can be called up according to any stored set of specifications, so that e.g. one could call up grammar items appropriate to some given ability range. It is also possible to specify limits e.g. for standard error of measurement, point biserial and content area. Among the advantages listed for this approach to testing are (i) that security problems are minimised, since the same form of the test need never be used more than once, (ii) that it is more efficient than traditional procedures, and (potentially) more reliable, and (iii) that it offers a less expensive way of equating different forms of a test, since new items can be added without recalibrating the whole set, simply by using a small, stable set of linking items. Henning also makes reference to the potential use of Rasch-based item banks in computer-administered adaptive testing:

> "From such a bank items can be drawn successively to match the ongoing performance of the examinee. A person who succeeds (or fails) with one item will be presented with a more (or less) difficult item in an iterative fashion until the actual ability of the examinee is located on the latent ability continuum." (Henning, 1986:73).

Theunissen (1987) is also concerned with the use of Rasch-based methods in computerised test design, and in particular with the banking of clusters of items based e.g. on the same reading comprehension text. A procedure for retrieving single items, based on optimisation theory, and using the test information function, is first described. An extension to this method for use with clusters of items is then suggested: in this case, the notion of individual item information is replaced with that of subtest information (i.e. the sum of the item information functions for the items in a given cluster). The subtest information functions are used in a similar way to the single item information functions in the simpler procedure, to select for retrieval from the bank the minimum number of texts required to meet a given test information specification.

De Jong (1983, 1984a, 1984b, 1986a, 1986b) describes studies in which Rasch-based methods were used in the development of measures of foreign language listening comprehension for purposes of national certification in the Netherlands.

In the first of these studies, Rasch fit statistics were used as the basis for the selection, from a larger item pool, of a set of items which would form a valid measure of listening comprehension (defined for these purposes as "the ability to understand the foreign language at the level of native speakers of comparable

age and educational background" (p.12)). Using the responses both of a target population group and of a suitable native speaker group, de Jong demonstrates the usefulness of a Rasch-based approach to construct validation by showing that the items which both (a) fitted the model, and (b) discriminated (in the expected direction) between the native and non-native speaking groups, could be considered to form a valid test of the construct of interest. Items which did not meet these requirements were deleted successively, and re-analyses carried out. Upon inspection, it appeared that those items depended on general intelligence, knowledge of the world and/or alertness rather than listening comprehension. Indeed, when analysed together they seemed to form a consistent measure, but a measure of something different from the ability measured by the 40 'best' listening comprehension items. De Jong (1983) draws the general conclusion that provided the majority of items in the try-out pool actually measure the intended ability, Rasch analysis can be of great help to the test constructor in identifying the items which are most likely to tap that ability.

De Jong (1984a) relates to the same investigation, but provides additional information concerning the two item formats used in this test (true-false and gap-filling, or listening cloze). When the two item subsets were analysed separately, it was found that 54% of the true-false items and 84% of the listening cloze items exhibited the required correspondence between native speaker response and the trait defined in the Rasch analysis of data from the target population. It was concluded that both item types could form the basis for valid and reliable measures of listening comprehension, but noted that the listening cloze items seemed to demonstrate better psychometric qualities.

In de Jong (1984b) the performance, on the same test, of 3 different groups of subjects was investigated. The groups consisted of: (i) 30 native speakers of English, aged approximately 17, and studying for 'A' levels, (ii) 44 native speakers aged 15-16, two years below American High School graduation level, and (iii) 575 non-native speakers aged 17, in their final year of secondary school (academic division) in the Netherlands. Again, the purpose of the investigation was to identify those items which together would form a valid and reliable measure of listening comprehension. Both traditional and Rasch analyses were conducted, firstly on data from the complete set of 59 items, and subsequently on data from the 40 'best' listening comprehension items and the 19 items identified as misfitting in the earlier study (de Jong, 1983). Using the complete item set, the results did not show the necessary differences among groups for this to be viewed as a satisfactory measure of listening comprehension. As in the earlier

study, however, it was concluded that the subset of 40 selected items constituted a valid measure, since they discriminated between two groups of native speakers of different ages and educational backgrounds, and, although there was overlap in the scores of the target E.F.L. group and the younger native speaker group, this was considerably less than when the 19 items thought to measure general intelligence or knowledge of the world were included in the test. The author points out that the English of pre-university level students in the Netherlands is of a fairly high level, as scores on the Test of English as a Foreign Language would confirm, and does not, therefore, consider this overlap with American High School students to be surprising. It is concluded from the results of the analyses presented in this study that the listening ability of native and non-native speakers can be measured along a single variable.

In the two other studies reported by de Jong (1986a, 1986b), Rasch-based methods were used for purposes of equating different tests of the same ability. In the first of these, the author illustrates a method used in determining the effectiveness of listening comprehension items (in English, French and German), at two different levels of ability, in circumstances where the items can be pretested only on a mixed ability sample and where the only information available is the estimated difference in mean ability and distribution of ability for the two different groups. (This situation arises as a result of there being a choice of levels at which students at a certain type of secondary school can be assessed in their final year.) Rasch statistics are used to determine whether or not the difficulty of each item is such that it is suitable for students within the relevant ability range, and if so, at which of the two levels it is more appropriate. The discriminating power of each item is also taken into account, by inspecting the steepness of the observed ICC, and this information used, together with the item difficulty estimates, in assessing the suitability of each item for measurement at the different levels. The effectiveness of the item selection method has been evaluated by comparing the predictions concerning the appropriateness of items with observed results using the final versions of the tests: the results have been found to be promising.

De Jong (1986b) is concerned with the monitoring of national educational standards, and focusses in particular on a method, based on Rasch analysis, for equating tests of different levels and from different years. The particular tests involved in this study are English and German listening comprehension tests, and the equating procedure is carried out by means of common 'linking' items, and common representative samples of testees. The purpose of the study was to

determine the effects of changes made in one sector of the secondary school system on the achievement of the pupils concerned. The conclusion drawn from the results, both on the listening tests and on E.F.L. reading comprehension tests for which results are also examined, was that the changes have lowered the achievement level. The Rasch-based method illustrated in the study is recommended by the author as a means of monitoring national standards over time.

3.1.3 Use of Rasch Analysis in the Investigation of Language Test Data

Two of the studies which use Rasch analysis to investigate various aspects of data obtained from language tests are concerned with bias in English language proficiency tests.

In the study by Chen and Henning (1985) data from the English as a Second Language Placement Examination used at U.C.L.A. were analysed, in order to investigate the nature, direction and extent of any bias that may be present for two linguistically and culturally distinct subgroups of examinees. Rasch analyses were performed on the responses of 34 native speakers of Spanish and 77 native speakers of Chinese (Mandarin or Cantonese dialects), both as separate groups and as a single group. The mean raw scores for the Chinese subgroup were higher than those for the Spanish subgroup on all subtests except for the vocabulary subtest: the authors suggest that the Spanish speakers might be expected to be favoured by this part of the test, in view of the morphological similarities between Spanish and English. The Pearson correlations computed for the sets of Rasch difficulty estimates obtained separately from the two groups indicated that the weakest relationship between difficulty estimates was to be found in the vocabulary subtest ($r = 0.31$). Those for the other subtests were 0.89, 0.73, 0.74 and 0.76. When the sets of estimates were plotted for all items, 4 items, all from the vocabulary subtest, were found to lie outside the 95% confidence interval constructed around the regression line, and all of these were biased in favour of the Spanish speaking group. On closer inspection, these items were found to hinge upon English words for which there are close cognate forms in Spanish but not in Chinese.

As the authors note, additional items would have been identified as biased if narrower confidence intervals had been set. They acknowledge, too, that the 'bias' they have identified could be viewed simply as a manifestation of the advantage which, as a result of lexical similarities, speakers of Spanish are likely to have over speakers of Chinese when learning English. They therefore take the

view that unfair bias, in a language testing context, could be said to exist if a disproportionate number of items favouring one group were included, i.e. a proportion exceeding that which occurs naturally in the language being tested.

The study by Madsen and Larson (1986) was also concerned primarily with bias relating to student language background. Before the main investigation was undertaken, two preliminary checks were carried out, using contrived and simulated bias, in order to ensure that the Rasch analytic procedures envisaged for use would indeed be capable of detecting bias. Since the results of these proved satisfactory, a larger, genuine investigation was carried out, using parts of an E.S.L. placement test battery, administered to 183 students ranging in levels from beginner to intermediate. 55% of the students were Spanish; the remainder were of a variety of nationalities and language backgrounds. Separate analyses of the different subtests were carried out, in order to avoid violating the assumption of unidimensionality.

The expected differences in fit among the different language groups were not found; there was, however, some evidence of greater misfit among low-ability students, particularly on the grammar and listening subtests. This suggested that these parts of the test had been too difficult for some of the lower-level students. As the authors note, erratic performance can be caused by many factors other than bias, and findings relating to bias are in any case specific to the particular tests studied.

The question of the applicability of the Rasch model to language proficiency measures consisting of several subtests intended to tap different skills is addressed in a study by Henning, Hudson and Turner (1985). In order to assess the robustness of the Rasch model to what might be expected to be violations of the assumption of unidimensionality, the authors performed Rasch analyses on the responses of 312 adult non-native speakers of English on the U.C.L.A. English as a Second Language Placement Examination. This test consists of subtests for listening comprehension, reading comprehension, grammar accuracy, vocabulary recognition and writing error detection, each containing 30 four-option multiple-choice items. Separate analyses were conducted on each subtest separately, and on the whole test. Following the procedure for checking for unidimensionality proposed by Bejar (1980), difficulty estimates obtained from the subtest calibrations were plotted against those from the whole-test calibration. The clustering of the points around a straight line of unit slope passing through the origin, observed in all 5 cases, was taken as indicating that the data

conformed to the unidimensionality assumption. Henning et al. report that no item points fell outside the 95% confidence intervals constructed around the regression lines, and that only one fell outside the 68% confidence interval. The conclusion that the unidimensionality assumption was not violated, despite the diversity of subtest content and sample characteristics, was supported by two further checks: the t-test of observed differences in difficulty estimates, and the comparison of the frequency of misfitting items in the subtest and the whole-test analyses.

It is in connection with this study by Henning et al. that Spurling (1987) raises doubts concerning the validity of Bejar's proposed method for investigating the dimensionality of data (see Section 2.3.6.4); as regards the other evidence for unidimensionality presented by Henning et al., he raises no objection, however.

3.2 Background to Applications in this Study

3.2.1 Issues for Investigation

As was indicated in Chapter 1, use of Rasch analysis represents a very recent development in the field of second/foreign language testing. It is clear from the reported studies described above, however, that its introduction has (potentially) profound implications for the design and development of language testing procedures. The reported applications also raise a number of interesting and important issues which, in view of the benefits which seem likely to be gained from the appropriate use of Rasch-based statistics, merit further investigation.

One of the first questions which comes to mind in considering the possible value of this approach in language testing is that of the types of language tests with which it might be used. The majority of applications reported to date have involved multiple-choice or true-false items rather than constructed response items. It would be of interest, therefore, to explore the use of Rasch analysis with other types of dichotomously-scored items.

An issue of particular concern to language testers is the extent to which data from language test batteries might be considered unidimensional in the sense required by the Rasch model. The familiarity of the division of language into subcomponents for purposes of teaching and testing is such that the notion of unidimensionality in language testing may seem untenable; indeed, some would view it as restrictive and undesirable.

Investigations of the fit of language test data to the Rasch model are

therefore necessary, with more detailed examination of the possible reasons for identified person and item misfit than have tended to be offered in published studies (with the exception of those of de Jong, in which examples of well-fitting and misfitting items are given). Checks on the extent to which guessing and time limit effects may affect test unidimensionality would also be of interest, since these are factors which can, at least to some extent, be controlled in language test design.

The sample-independence of difficulty estimates and the test-independence of ability estimates are seen as being major advantages offered by the Rasch approach when fit between model and data is sufficiently good. Investigations of the circumstances under which the desired stability of estimates is achieved are needed, however: it cannot simply be assumed that this feature will obtain, though this is the impression sometimes given.

The view of items as appearing in the same order of difficulty for all members of the target population is one which may meet with scepticism from language teachers and testers alike, since it runs counter to the widely-held idea that students of different educational and linguistic backgrounds experience different problems in learning English. In the light of approaches to second language acquisition in which it is suggested that acquisition might follow a consistent developmental sequence, however, the notion of a consistent ranking by difficulty is perhaps less implausible than in relation to knowledge of other types. The results of investigations of this matter, as well as being important from the point of view of language testing, would also be of interest to those engaged in research into the development of second language ability.

Notwithstanding the comments of Hambleton and Murray (1983), quoted earlier, it would also be useful to carry out further comparisons of the information obtained using traditional and Rasch approaches to test analysis: as has been seen, the criteria for comparison, and the relationships between traditional and Rasch indices, have not always been made clear, with the result that reasons given for preferring a particular method sometimes seem spurious.

3.2.2 Background to Test-Types Used

Details of the composition of the two English proficiency tests analysed in this study are given later; it is appropriate here, however, to note the main differences between them in terms of the approaches to proficiency which they embody.

As Stern (1983) observes, various attempts have been made to conceptualise and describe proficiency in a second language. He notes that in many of these, proficiency has been defined in terms of a number of psychological and/or linguistic components: the particular categories used have varied, but the general view of language proficiency as being divisible into a set of component parts has been maintained. An alternative view, that of language proficiency as being essentially unitary (a standpoint associated in particular with John Oller) has, as Stern remarks, challenged these other definitions.

This second view is essentially that which underlies the cloze-type test discussed in Chapter 4. This test seeks to measure proficiency in a global way, by means of a single technique. No attempt is made, in the design of such a test, to specify the content of each item: it is assumed that the task of restoring deleted words to text (or, in this case, finding acceptable substitutes for them) draws upon some overall language ability.

There have, of course, been studies which have sought to identify the particular abilities required by tests of this kind (see e.g. Alderson, 1978, 1979, 1980; Bachman, 1982, 1985; Lee, 1985). However, unless some form of rational deletion of words is used, based on item content, construction of such tests is usually carried out without regard to specific abilities.

The analyses presented in Chapter 5, on the other hand, are based on data from three subtests of a language test battery (the ELTS test). As Criper and Davies (1986:10-11) explain, the construct of language proficiency embodied in this test includes divisions on three dimensions: (a) a skills dimension (reading, listening, writing, speaking), (b) a general vs study dimension, and (c) a specialist subject dimension. Within each subtest, items are further subdivided according to Munby's (1978) taxonomy of skills and micro-skills.

Thus the two tests analysed in the following chapters may be seen as representing opposite extremes in terms of the degree to which item content is specified. In view of this difference, one might expect data from the cloze type test to conform more closely to the Rasch assumption of unidimensionality than the ELTS data. Some of the investigations reported in Chapters 4 and 5 will be concerned with the dimensionality of the two data sets.

CHAPTER 4
ANALYSIS OF CLOZE-TYPE TEST DATA

In this chapter, the results of traditional and Rasch analyses of response data from a cloze-type test are first presented and compared. Attention is then focussed on the extent to which these data might be considered to meet the requirements of Rasch measurement, and on the question of whether the advantages offered by the use of Rasch analysis are realised in this application.

4.1 Description of the Cloze-Type Test Data

The cloze-type test used in this study was constructed by Dr. Clive Criper for the Ministry of Education in Malaysia in 1975. It was designed both for purposes of placing learners of English on a graded reading scheme (the Edinburgh Project in Extensive Reading), and for use as a measure of the general English proficiency of school pupils and teachers of English in Malaysia, at levels ranging from near-beginner to advanced. The test has also been used as a placement test for students entering general courses in English as a Foreign Language at the Institute for Applied Language Studies, University of Edinburgh.

4.1.1 Composition of the Test

The test consists of 12 separate passages (average length 80 words), taken from graded readers of various levels, and arranged in order of increasing difficulty. Each passage contains between 10 and 15 blanks: the deletions were selected initially by leaving a 9-word introduction to each passage, and then deleting every 6th word thereafter. However, some of the items created by this procedure were found on the basis of pretesting to be unsatisfactory in terms of difficulty level or discriminating power, or to be problematic in some way, e.g. requiring the restoration of proper nouns, or having too many possible answers to be scored efficiently and reliably. The strict deletion pattern was therefore modified in such cases, by moving the blank one word to the left or right, or, in a very small number of cases, by dispensing with it altogether.

The final version of the test, which has 141 deletions, was arrived at after a repeated process of editing and piloting, involving a total of 100,000 pupils and teachers from Upper Primary and Secondary Schools in Malaysia. It should be noted that the final deletion pattern arose out of the attempt to avoid creating items which might result in inefficiency or inconvenience, and was not

deliberately chosen in order to include particular word classes or item types.

A copy of this cloze-type test, together with the marking scheme, appears in Appendix B; although the test paper has been reduced in size for inclusion here, the format of the passages, including the positions of the line breaks, has been retained. The extent to which modifications to the strict n^{th} word deletion pattern proved necessary can be seen from Table 4.1 below, where the deletion pattern in each of the 12 passages is shown.

Passage label	*Deletion pattern	Words in passage	No. of blanks
A	9_5_5_5_4_5_8_4_5_5_5_2	73	11
B	11_5_6_4_5_5_5_5_5_6_6_4_7_1	88	13
C	9_5_5_4_5_5_5_6_5_5_7_5_5_1	85	13
D	9_5_6_6_5_6_4_5_5_5_4_5_5_6_4_2	97	15
E	8_5_6_5_11_5_5_5_5_5_4_5_	81	12
F	9_4_6_5_5_5_5_7_6_6_7_5_3	85	12
G	9_5_5_5_5_5_5_5_5_5_5_5_5	75	11
H	9_5_11_5_5_5_5_5_5_5_5_1	71	10
I	8_4_4_4_6_5_4_6_5_5_5_7	74	11
J	9_8_6_4_5_5_5_5_5_5_5_6	79	11
K	9_5_5_5_5_5_6_4_6_5_	65	10
L	10_4_11_6_6_5_5_6_4_5_7_5_2	88	12
		141	

* '_' = deleted word
Numbers refer to words remaining after deletion.

Table 4.1 Deletion Pattern in the Cloze-Type Test

The composition of the test in terms of the form classes of the deleted words is shown in Table 4.2 below. The figures in brackets are the corresponding percentages for the test as it would have been without modifications to the original deletion pattern.

Word class	%	
Main verb	19.9	(18.4)
Noun	12.1	(22.5)
Adverb	5.7	(6.1)
Adjective	7.8	(7.5)
TOTAL CONTENT WORDS	45.5	(54.5)
Auxiliary verb	3.5	(5.4)
Verb particle	3.5	(2.0)
Adverbial particle	0.7	(-)
Pronoun	15.6	(10.9)
Determiner	10.6	(9.5)
Preposition	15.6	(13.6)
Conjunction	4.3	(3.4)
Negative particle	0.7	(0.7)
TOTAL STRUCTURE WORDS	54.5	(45.5)

Table 4.2 Form Classes of Deleted Words in the Cloze-Type Test

Comparison of the two columns in Table 4.2 above indicates that only the noun and pronoun categories changed by more than 2% as a result of the modifications, with the greatest difference being in the proportion of nouns. The need for changes to items in the noun category will have been contributed to by the fact that 5 of the blanks in the initial version of the test fell on proper nouns, which would indeed have been expected to perform unsatisfactorily as test items.

For the purposes of this discussion, each item in the test is identified by a letter (A - L) denoting the passage to which it belongs, and by a number (1 - 141) indicating its position in the test as a whole.

4.1.2 Administration and Scoring

The time limit of 1 hour set for this test is intended to allow testees to attempt all passages; however, in view of the graded design of the test, lower-level candidates are not expected to be able to complete all of the blanks.

Candidates are instructed to supply a single word for each blank, and are shown a short example passage, with answers, before beginning the test. They write their answers on a separate answer sheet containing the numbers 1 to 141, with a blank space for each answer.

The scoring procedure used is a form of acceptable word marking, using a prepared answer sheet containing the words suggested by a group of native speakers. The number of possible fillers specified for each blank ranges from 1

to 8, and, notwithstanding the editing process described above, there are 5 cases in which the scorer is instructed to mark as correct any acceptable filler, since the possible answers were too numerous to list. Answers are marked correct only if spelt correctly.

4.1.3 Description of Samples

The larger of the two data sets analysed in this chapter consists of the responses of 611 Malaysian testees drawn from the total sample of 100,000 learners tested during 1976-77 in the course of developmental work on the test. This group of 611 was selected so as to contain learners of as wide a range as possible of the proficiency levels spanned by the test.

The second data set consists of the responses of 243 Tanzanian learners tested in Tanzania in 1984 as part of an investigation of the teaching and learning of English in Tanzania (see Criper & Dodd, 1984). This group, which was also drawn from a larger testee sample (consisting mainly of Primary Six and Secondary School pupils, but including some teachers of English), was again selected so as to contain learners of all the various levels found in the total sample. This second data set was used primarily for purposes of comparison with the main set.

Data collection was in both cases carried out either by, or under the direction of, Dr. Criper. Both the Malaysian and Tanzanian samples used in this study were selected solely for purposes of test analysis, and not in order to be representative of any particular population. However, since the test was intended to serve as a measure of general proficiency for learners of English of any background, and not only for those in Malaysia, both samples can be seen as belonging to the (wider) target population for the test. Furthermore, since rigorous sample selection procedures are sometimes not practicable in test development work, it would not be unusual for samples such as these to be used for the pretesting of items.

For both groups, scoring of the answer papers was done by class teachers. For the purpose of the analyses carried out here, the scored responses were coded as correct, incorrect or omitted. Although the two methods of analysis used make use only of the first two categories (omitted responses in both cases being counted as incorrect), the record of omissions will be referred to in considering matters such as the possible effects of the time limit.

For this study, analyses were performed both using the complete data sets as described above, and using subsets drawn from these in various ways, e.g. on the basis of item content, or of scores obtained. Details of the particular subsets used are given in the relevant sections of the discussion.

4.2 Traditional Analysis of Cloze-Type Data

This section is concerned with the information yielded by traditional analyses of the two basic sets of cloze-type data referred to above. The results for the Malaysian and Tanzanian groups are compared, since it is of interest to determine the extent to which, using this method of analysis, the information obtained about the test is similar for two different subpopulations drawn from the target population for the test.

4.2.1 Traditional Statistics Computed

For each of the two data sets, the following traditional statistics were computed:

1. The total raw score for each person;
2. The facility value (proportion correct) for each item;
3. The E_{1-3} discrimination index for each item, using subgroups of 27% of the total sample;
4. The 'unbiased' point biserial correlation coefficient for each item (i.e. the correlation between dichotomous responses and total test scores, with the item in question being removed from the total score in each case);
5. The K-R20 estimate of internal consistency reliability for the set of responses;
6. The standard error of measurement for the set of scores.

4.2.2 Summary and Interpretation of Results

The results of these analyses are given in Appendices C and D, with the exception of the raw scores, which, in view of the large number of persons involved, are not listed individually, but are instead summarised in the form of frequency counts and histograms (see Appendices C.1 and D.1).

The individual item statistics are set out in Appendix C.2 (for the Malaysian group) and Appendix D.2 (for the Tanzanian group). These are summarised and re-ordered, for ease of interpretation, in Appendices C.3 and D.3. Each of these two appendices contains 3 tables, in which the items are grouped according to

the intervals in which they fall on the 3 different indices calculated. These tables provide a general indication of the distributions of the item statistics for the data set in question, and allow ready identification of items at the extremes of the scales.

4.2.2.1 Raw Score Distributions

The information contained in Appendices C.1 and D.1 is further summarised in Tables 4.3 and 4.4 below, to provide a general indication of the raw score distributions for the Malaysian and Tanzanian groups.

Raw score range	Frequency count	% of group	Cumulative %
0 - 19	28	4.6	4.6
20 - 39	48	7.8	12.4
40 - 59	54	8.8	21.2
60 - 79	100	16.4	37.6
80 - 99	116	19.0	56.6
100 - 119	157	25.7	82.3
120 - 141	108	17.7	100.0
	N = 611		

Raw score range = 0 - 136
Mean raw score = 86.2
SD of raw scores = 33.4

Table 4.3 Raw Scores Obtained by Malaysian Group on Cloze-Type Test

Raw score range	Frequency count	% of group	Cumulative %
0 - 19	34	14.0	14.0
20 - 39	44	18.1	32.1
40 - 59	43	17.7	49.8
60 - 79	46	18.9	68.7
80 - 99	44	18.1	86.8
100 - 119	31	12.8	99.6
120 - 141	1	0.4	100.0
	N = 243		

Raw score range = 3 - 129
Mean raw score = 59.3
SD of raw scores = 32.2

Table 4.4 Raw Scores Obtained by Tanzanian Group on Cloze-Type Test

Comparison of Tables 4.3 and 4.4 above indicates that the Tanzanian group is in general lower in proficiency (as measured by this test) than the Malaysian group: almost 50% of the Tanzanian testees obtained raw scores of less than 60, compared with only about 20% of the Malaysian group, and the mean scores for the two groups differ by approximately 27 raw score points. Although the standard deviations are very similar, the raw scores for the Malaysians extend higher in the possible range (highest score = 136, as opposed to 129), and this group contains a considerably larger proportion of persons scoring 100 or more.

4.2.2.2 Item Facility Values

When calculated from the responses of the Malaysian testees, the facility values for the 141 items in this test have a mean of 0.61 (SD=0.23), and are distributed as shown in Figure 4.1 below.

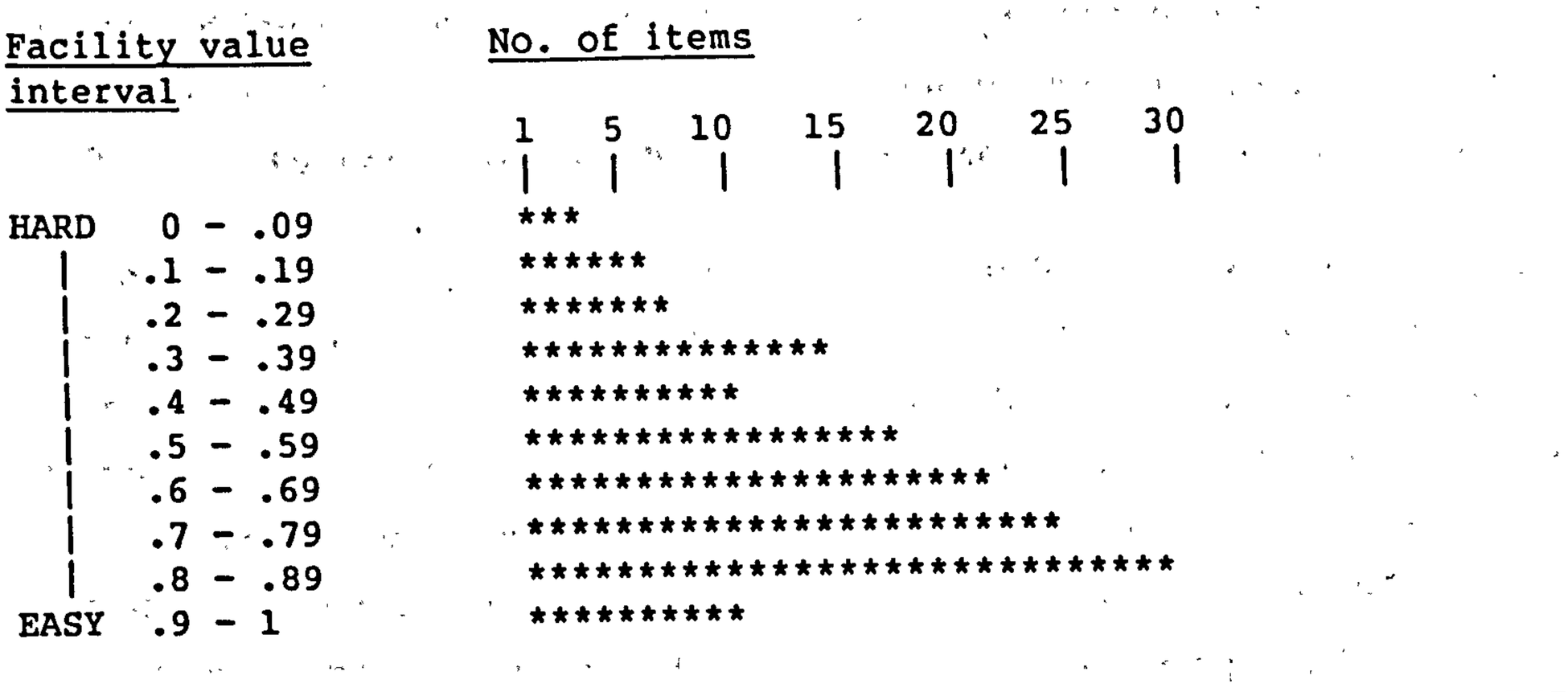


Figure 4.1 Distribution of Facility Values for Cloze-Type Test (Malaysian Data)

For the Malaysian group, over 70% of the items fall above the midpoint of the facility value scale: only 40 of the 141 items have values of less than 0.5. Comparison of Figure 4.1 above with the corresponding distribution for the 243 Tanzanian learners (shown in Figure 4.2 below) indicates that for the Tanzanian group, there are more items at the extreme of difficulty, and fewer at the extreme of easiness.

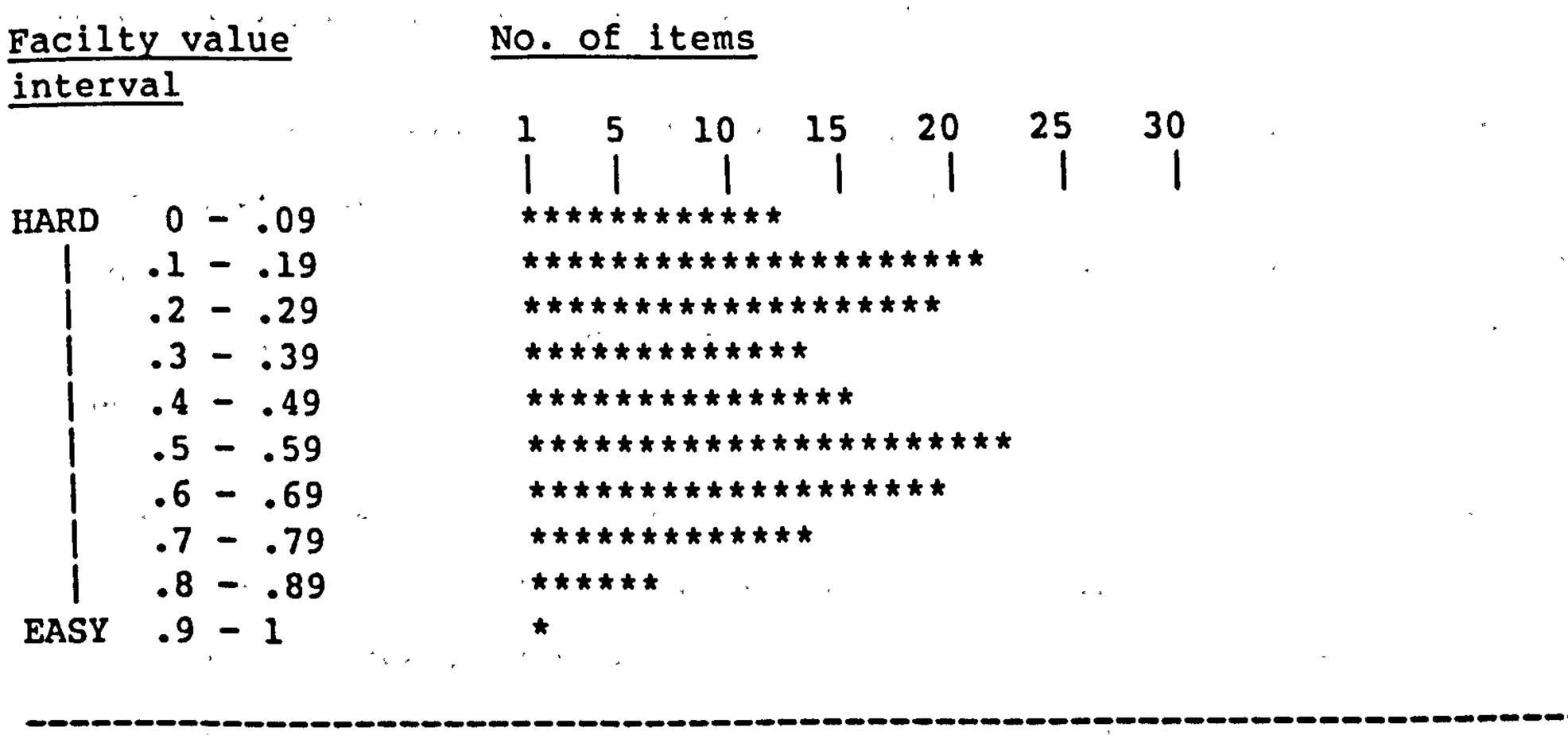


Figure 4.2 Distribution of Facility Values for Cloze-Type Test (Tanzanian Data)

For the Tanzanian group, the mean facility value is 0.42 (SD=0.24). Only about 43% of the values fall at or above the midpoint, leaving 80 items with values of less than 0.5, thereby giving the impression of a generally more difficult test. Thus if, as is frequently the case in test development, the purpose in calculating

the facility values had been to identify, and then to discard or modify, items falling outside some chosen limits (e.g. .33 - .67, as in the 'rule of thumb' quoted by Henning, 1987:50), then the changes made to this test would have differed quite markedly, depending on which of the two sets of data had been used.

In this case, of course, such a procedure would not have been appropriate, since, quite apart from the problems inherent in manipulating the difficulty levels of cloze-type items, the requirement here was for a test which could separate testees into approximately 10 different proficiency bands, and in which, therefore, item difficulties were fairly uniformly distributed throughout the range. Again, though, the improvements which would be recommended on the basis of these results would not be the same for the two data sets: using the Malaysian data, the deficiency appears to be largely in the middle-to-difficult portion of the scale, while the results for the Tanzanian group suggest that more very easy items/passages are required.

As regards the easiest and most difficult items identified in the two separate analyses, comparison of Table 1 in Appendix C.3 with Table 1 in Appendix D.3 indicates that these are similar for the two testee groups: the 10 easiest items identified in each analysis contain 7 items in common, and the 10 most difficult items contain 6 items in common. Indeed, there is quite close correspondence between the two complete sets of facility values, as is indicated by the Pearson product-moment correlation coefficient of 0.85 ($p < 0.001$).

Comparison of the pairs of facility values themselves, however, reveals, for the complete set of items, an average magnitude of difference of 0.2, with differences ranging from 0 to 0.73. Although there are 28 items for which the two values differ by less than 0.1, there are 62 for which the difference is greater than 0.2.

Thus although the facility values for the two testee groups show a fairly close relationship in terms of their relative distances, in standard deviation units, from their respective means, the pairs of values themselves are not the same, and the impression of the overall difficulty of the test, and of the extent to which the intended design seems to have been achieved, differs somewhat for the two data sets. These differences, of course, result from the differences in the distributions of proficiency for the two groups, referred to in the previous section.

4.2.2.3 Indices of Discrimination

The E_{1-3} discrimination statistics for the Malaysian data set range from 0.09 to 0.93, with a mean of 0.58. As is obvious from the minimum value, none of the items has resulted in reverse discrimination. It can be seen from Table 2 in Appendix C.3 that the least effective discriminators identified by this index are the 7 items with values of less than 0.2, i.e. those on which the success rates for the high- and low-scoring subgroups differed by less than 20%. The subgroup size for this analysis is 165 (i.e. 27% of the total sample size).

Decisions concerning minimum acceptable levels of discriminating power must, of course, depend (a) on the distribution of proficiency within the person sample, since values on this index would be expected to be higher for a heterogeneous group than for a homogeneous one, and (b) on the required distribution of item difficulties within the test, since some very easy and/or very difficult items, though ineffective for purposes of discrimination, might need to be retained for reasons of overall test design.

As the high- and low-scoring subgroups in the Malaysian sample are known in advance to differ widely in proficiency levels (the sample having been chosen so as to ensure this), the discrimination indices should in this case be generally high, though, in view of the effort made to include items of a variety of difficulty levels, not uniformly so. At the same time, one would not wish to reject outright items whose low discriminating power results entirely from extreme easiness or difficulty, since some such items are necessary to the design of the test. Thus for the purposes of the test under discussion here, the discrimination statistics must be interpreted in conjunction with the item difficulty statistics: viewed alone, they provide insufficient information on which to base judgements about the performance of individual items.

Of the 7 items identified in this analysis as being the least effective discriminators, 3 are of extreme easiness (facility values $\geqq 0.94$) and 2 are of extreme difficulty (facility values $\leqq 0.08$), leaving only 2 (items B13 and C34, with discrimination indices of 0.09 and 0.16) whose low discriminating power may be attributable to other factors, and whose performance might therefore be questioned. The facility values for these are 0.77 and 0.81 respectively.

Item B13 appears in the second sentence of the second passage, in the following context: 'Jenny Lim and her brother Peter went for a walk. As (B12) passed the big house on (B13) hill a dog ran out.' The only answer accepted as

correct was 'the'. Success rates on this item for the high- and low-scoring subgroups were 85% and 76% respectively. A brief examination of the actual answers given by the Malaysian testees reveals that many of those who 'failed' on this item had supplied the indefinite article instead of the definite article, and that this choice of answer was made by members of both subgroups.

Although the definite article might in this case be the automatic choice for most native speakers, the indefinite article nevertheless also seems acceptable in the context. Had both of these answers been counted as correct, this item would no doubt have shown low discrimination as a result of extreme easiness rather than for any other reason. It would appear, then, that any perceived inconsistency in the functioning of item B13 in the test as it stands stems from failure to anticipate, and hence to include in the marking scheme, this alternative possible answer.

The context for item C34, which appears towards the end of the third passage, is as follows: '"... I've brought you (C34) fruit. I said to myself, 'I must (C35) Mrs. Chong some of my (C36)' ..."'. Success rates for the high- and low-scoring subgroups were 81% and 64% respectively. For an item with a pass rate of 81% for the whole sample, one might have expected an even higher pass rate among the upper group. Looking at the answers supplied for this item, one finds that of those who did not give the correct answer ('some'), many supplied 'a'. In order to understand the behaviour of this item, it would be of interest to know how the testees arrived at their answers; it is, unfortunately, not possible in the present study, to interview the individuals concerned. It may be, however, that in this case the choice of the indefinite article was prompted by the (ostensibly) singular form of the noun, and that some of the higher level testees were unduly influenced by considerations of formal agreement.

Comparison of Table 2 in Appendix C.3 with Table 2 in Appendix D.3 shows that the E_{1-3} discrimination statistics calculated from the Tanzanian data are distributed somewhat differently from those for the Malaysian group: the values for the Tanzanian group range from -0.02 to 0.98, with a mean of 0.57. (Subgroup size was in this case 66.)

The difference in the observed distributions, particularly at the lower end of the scale, provides an illustration of the difficulty of setting a minimum acceptable value for the discrimination index, and hence of the need to interpret discrimination levels in a relative way. As with the two sets of facility values, this difference can be attributed to the difference between the Malaysian and

Tanzanian samples in terms of the distribution of proficiency within each.

The least effective discriminators identified using the Tanzanian data set are the 8 items with values of less than 0.1 (see Table 2, Appendix D.3). Of these, 7 are of extreme difficulty for this group (facility values ≦0.04), and thus could not be expected to differentiate between the high- and low-scoring subgroups. The remaining item (B13), on the other hand, has a facility value of 0.76, and, furthermore, is the only one which discriminates in reverse (discrimination index = -0.02). The success rate on this item was 74% for the high scorers and 76% for the low scorers. Examination of the answer papers again shows that for those who 'failed' on this item, the indefinite article was a common choice of answer, and that, in the case of the Tanzanian group, it was supplied even more frequently by the higher-level testees than by the lower-level ones. Possible reasons for this might be that the less proficient testees tended to choose 'the' by analogy with the definite article occurring earlier in the sentence, while some of those who were more proficient may have looked ahead in the sentence and been influenced by the indefinite article occurring later; it is also conceivable that some oversimplified, taught 'rule' concerning the use of the indefinite article for the first mention of a referent may have come into play in some cases. Again, though, the apparent problem with this item could be cured simply by adding the indefinite article to the marking scheme.

It will be noted that the same item, B13, has the lowest discrimination index for both the Malaysian and the Tanzanian groups, and is identified in both analyses as behaving suspiciously rather than failing to discriminate because of its difficulty level. The second item shown by the Malaysian results to be questionable, item C34, does not figure among the poor discriminators in the Tanzanian results at all, however: indeed, with a facility value of 0.51 and a discrimination index of 0.79, it would, by traditional criteria, be considered almost a model item.

In the 8 poorest discriminators identified for each group, there are only 3 common items: item B13, plus two others (I98 and J111) which were answered incorrectly by almost all testees in both groups. The differences in the remaining items identified result from differences in the proficiency levels of the groups: those which are of low discrimination for the Tanzanian group because of extreme difficulty show higher discrimination for the Malaysian testees, for whom they proved less difficult, and those which are of low discrimination for the Malaysian group because of extreme easiness show higher discrimination for the

Tanzanian group, for whom they proved less easy. Comparison of the results for the two groups, then, illustrates the influence of characteristics of the person sample on the information obtained from this type of analysis.

The other traditional index of item discriminating power calculated here, the (unbiased) point biserial, also reflects the difference between the two person samples, as can be seen by comparing Table .3 in Appendix C.3 with Table 3 in Appendix D.3. Although the mean values for the two groups are almost identical, the range of values for the Tanzanian group again extends lower than that for the Malaysian group, the lowest values being -0.01 and 0.1 respectively. The 7 poorest discriminators identified for each group by this index again contain only 3 items in common (the same items as for the E_{1-3} index). Where they do not correspond, this again seems to be largely attributable to the difference between the two groups in terms of levels: the additional items identified in the Tanzanian analysis were all of extreme difficulty (facility values $\leqq$0.04) for that group, but less difficult for the Malaysian group, for whom they showed better discrimination.

Of the additional items identified in the Malaysian analysis, one (A1) was extremely easy for the Malaysian group but slightly less so for the Tanzanian group. For the other 3 (C34, F73 and L136), though, extreme easiness or difficulty does not seem to provide a complete explanation. Although these all appear quite close to one or other extreme of the facility scale for the Malaysian group, there are other, more extreme items with higher point biserial coefficients. Given that these 3 items appear among the poorest discriminators for this group, they might be thought to merit further investigation. It should be noted, however, that with point biserials from 0.27 to 0.29, these items would frequently be considered adequate in terms of discrimination; although the need for flexibility in setting a minimum level is sometimes mentioned (see e.g. Thorndike, 1982b:26; Henning, 1987:53), lack of time or resources may in practice often mean that a minimum value of .25 or even .2 is used as a matter of routine.

For both of the traditional indices of discrimination used here, product-moment correlation coefficients were calculated for the values obtained for all 141 items from the two different testee groups. For the E_{1-3} indices (Malaysian vs Tanzanian group), $r = 0.27$ ($p < 0.001$), while for the point biserials, $r = 0.57$ ($p < 0.001$), indicating that although the point biserials show greater consistency between groups, they are nevertheless subject to some degree of sample-dependence.

As regards the 7 least discriminating items identified by the two different indices for the same testee group, it is found that for the Tanzanian group, these correspond exactly. For the Malaysian group there is also close correspondence, with 5 of the 7 appearing in both lists. Thus although, for the complete test, the point biserials obtained from the two groups show a closer relationship than the E_{1-3} discrimination statistics, there is little difference between the two indices in the particular items identified as being at the low extreme for the same group. Indeed, the two indices correspond closely across the whole set of items for the Malaysian group ($r = 0.82$), and extremely closely for the Tanzanian group ($r = 0.96$; $p < 0.001$ in both cases).

The results presented in this section, as well as providing some comparative information on the two indices, illustrate the need to take account of the relationship between discriminating power and difficulty when interpreting the traditional discrimination statistics. They also serve to demonstrate the sample-dependent nature of these statistics, and the inadvisability of using 'rules of thumb' unless the sample is known to be representative.

A final point which should be noted is that under the traditional approach to item analysis, attention is given only to items whose power to discriminate appears low. No maximum level for discrimination is set: indeed, the higher the value, the more effective and desirable the item is generally considered to be.

4.2.2.4 Test Reliability and Error of Measurement

The K-R20 estimate of internal consistency reliability is 0.98 for each of the two data sets. As was mentioned in Chapter 2, the estimation of reliability is related to certain characteristics of the test and of the person sample in question. Factors which are likely to have influenced these coefficients in a positive direction are (a) that the cloze-type test contains a large number of items, (b) that these are fairly homogeneous in content, and therefore intercorrelate generally highly, and (c) that the two groups of persons are both heterogeneous in terms of proficiency levels. A factor which is likely to have operated in the opposite direction is that the items in this test vary widely in difficulty.

The standard error of measurement is approximately 4 for both data sets. Since this is calculated from the reliability coefficient and the standard deviation of the raw scores, the relatively low values observed here reflect the high variance in the test scores obtained for both groups.

The fact that the indicators mentioned here are reported as single, global values for the whole data set implies that consistency of measurement will be the same for all persons in the sample, and that measurement error will be of similar magnitude at all points on the raw score scale.

4.3 Rasch Analysis of Cloze-Type Data

In this section, the results of Rasch analyses of the Malaysian and Tanzanian sets of cloze-type test data are summarised and interpreted.

4.3.1 Rasch Statistics Computed

For each of the two data sets, the Rasch statistics listed below were computed. Notes on the methods of calculation of these can be found in the specified sections of Appendix A.

1. The ability estimate corresponding to each raw score (except for 0 and 141), and its associated standard error (Appendices A.1 & A.2);
2. The (information-weighted) total fit t-statistic for each person (Appendix A.3);
3. The difficulty estimate for each item, and its associated standard error (Appendices A.1 & A.2);
4. The observed item characteristic curve for each item, across 6 roughly even-sized raw-score groups, and its proportional departure from model expectation (Appendix A.4);
5. The (information-weighted) total fit t-statistic for each item (Appendix A.3);
6. The between-group fit t-statistic for each item across 6 ability subgroups (Appendix A.4);
7. A Rasch model-based discrimination index for each item (Appendix A.5);
8. The person separability index for the data set (Appendix A.6);
9. The number of 'strata' into which the testees are separated by the test (Appendix A.6).

All the Rasch statistics except for no. 9 above were computed using BICAL (Wright, Mead & Bell, 1980) [1]. The estimation procedure used was UCON.

4.3.2 Summary and Interpretation of Results

The results of the Rasch analyses are set out in Appendix E for the Malaysian data set, and Appendix F for the Tanzanian data set.

Again, individual person statistics are not listed, because of the large number

of persons involved; they are, however, summarised in the form of frequency counts (in Appendices E.1 and F.1) and plots of fit against ability (in Section 4.3.2.2).

3 of the 611 Malaysian testees scored zero, and could not, therefore, be included in the Rasch analysis, for the reasons mentioned in Chapter 2. Thus for the purposes of this section, the Malaysian data set consisted of the responses of the 608 measurable persons remaining. No preliminary editing of the Tanzanian data set was necessary, there being no scores of zero or full marks, and so the total sample size in this case was, as in the previous section, 243.

4.3.2.1 Person Ability Estimates

The Rasch ability estimates corresponding to each raw score from 1 to 140, calculated using the Malaysian data set, are listed in the raw score-to-ability conversion tables in Appendix E.1. Two such tables are shown, one resulting from the initial analysis, based on the responses of all 608 persons, and the other resulting from the re-calibration carried out after the removal from the data set of the responses of 6 persons identified in the analysis of person fit as being 'misfitters' (see Section 4.3.2.2 below for details of this).

Comparison of these two tables shows that the effect on the ability estimates of omitting these persons' response data has been negligible: for 137 of the 140 points on the raw score scale, the difference between the two ability estimates corresponding to the same raw score is 0.02 logits or less, and in no case does it exceed 0.03 logits. Thus for this data set, it would make little difference which set of estimates was selected for use. However, since it is the second set (i.e. the estimates obtained after the removal of misfitting persons) which represents the final outcome of this method of analysis, and which would be used in any practical application, it is this set which will be referred to in this section. The results discussed in this part of the study are therefore based on a sample of 602 persons.

From Table 2 in Appendix E.1, then, it can be seen that, as far as the estimation of ability is concerned, the outcome of the Rasch analysis of the Malaysian data set is an ability scale ranging from -6.17 to 6.50 logits. The ability estimates for the individuals in the Malaysian group range from -6.17 to 4.70, these being the ability scale values corresponding to raw scores of 1 and 136 respectively. The mean ability for this group is 0.83 logits (SD = 1.79).

Each ability estimate in the table is shown with its associated standard error. It can be seen that that standard errors are lowest (≦0.22) for persons scoring between 41 and 92 on the raw score scale. As scores increase above this interval, or decrease below it, the standard errors increase in magnitude, gradually at first, but more markedly as scores approach the extremes for the test. As was indicated in Chapter 2, information is greatest, and hence the standard error lowest, for persons whose abilities are close in level to as many as possible of the items, i.e. for those near the centre of the possible ability range. For the ability estimates corresponding to the highest and lowest possible raw scores, the standard errors are 1.03 and 1.01 respectively, reflecting the relative lack of information that would be available in the response vectors of persons falling at the extremes of the range.

As the frequency count column in the table shows, not all possible raw scores were observed in the data set analysed here; of the 602 persons in the group, only one person (the person at the low extreme) has a score for which the standard error is greater than 0.5.

The raw score-to-ability conversion tables for the Tanzanian data set are shown in Appendix F.1. As before, two tables are given: the first resulting from the initial calibration, and the second based on the re-calibration after discarding the response data of misfitting persons (see Section 4.3.2.2).

The effect of omitting the misfitting persons (in this case 7), though greater than for the Malaysian data set, is again slight: comparison of the two sets of ability estimates shows that for 121 of the possible raw scores, the pairs of estimates differ by no more than 0.03 logits, and that the largest difference, observed at 4 points at the upper extreme of the raw score scale, is 0.06 logits. The slightly greater changes observed in the two sets of estimates for the Tanzanian data, as compared with those noted for the Malaysian data, are, at least in part, attributable to the difference in the proportions of persons omitted on grounds of misfit: for the Tanzanian group this is almost 3% of the sample, while for the Malaysian group it is less than 1%. (An additional possible contributory factor would be the actual degree of misfit shown by the persons concerned, since the more extreme the misfit, the more serious the disturbance to the results of the analysis.)

The final ability scale, constructed using the response data of the remaining 236 Tanzanian testees, ranges from -6.27 to 7.09 logits (see Table 2 in Appendix F.1). The ability estimates for the persons in this group range from -5.12 to 3.63,

these being the estimates corresponding to raw scores of 3 and 129 respectively. The mean ability for the group, −0.74 (SD = 1.78), is 1.57 logits lower than that for the Malaysian group, reflecting the general difference in the levels of the two groups.

From the standard errors listed against the ability estimates in this table, it can again be seen that the most confident ability estimates have been made for persons in the middle of the range: the standard errors are lowest (0.22) for the estimates corresponding to raw scores of 43 to 89. The largest standard error for any score actually observed in this data set is 0.6 (for the person with a raw score of 3). At the upper and lower extremes of the complete ability range for the test, the standard errors would be 1.09 and 1.02 respectively.

Figure 4.3 below shows each possible raw score plotted against its Rasch ability scale equivalent (i.e. the 'test characteristic curve') for each of the two data sets; each raw score and its Rasch ability equivalent from each of the two final conversion tables referred to above is represented in this figure by a single point. Those from the Malaysian analysis are represented by circles and those from the Tanzanian analysis by triangles. Also shown are the ranges of ability for the two testee groups, and the group means.

As is clear from Figure 4.3, the points plotted for the two different samples correspond extremely closely; indeed, in the middle to upper portion of the scale they coincide. The increasing distances between them for ability estimates approaching the extremes reflect the relatively large standard errors associated with these. Some examples of the pairs of values plotted in Figure 4.3, taken from throughout the score range, are set out for comparison in Table 4.5 below, together with their standard errors. It can be seen from these that taking into account the standard errors in each case, the pairs of ability estimates can be considered equivalent.

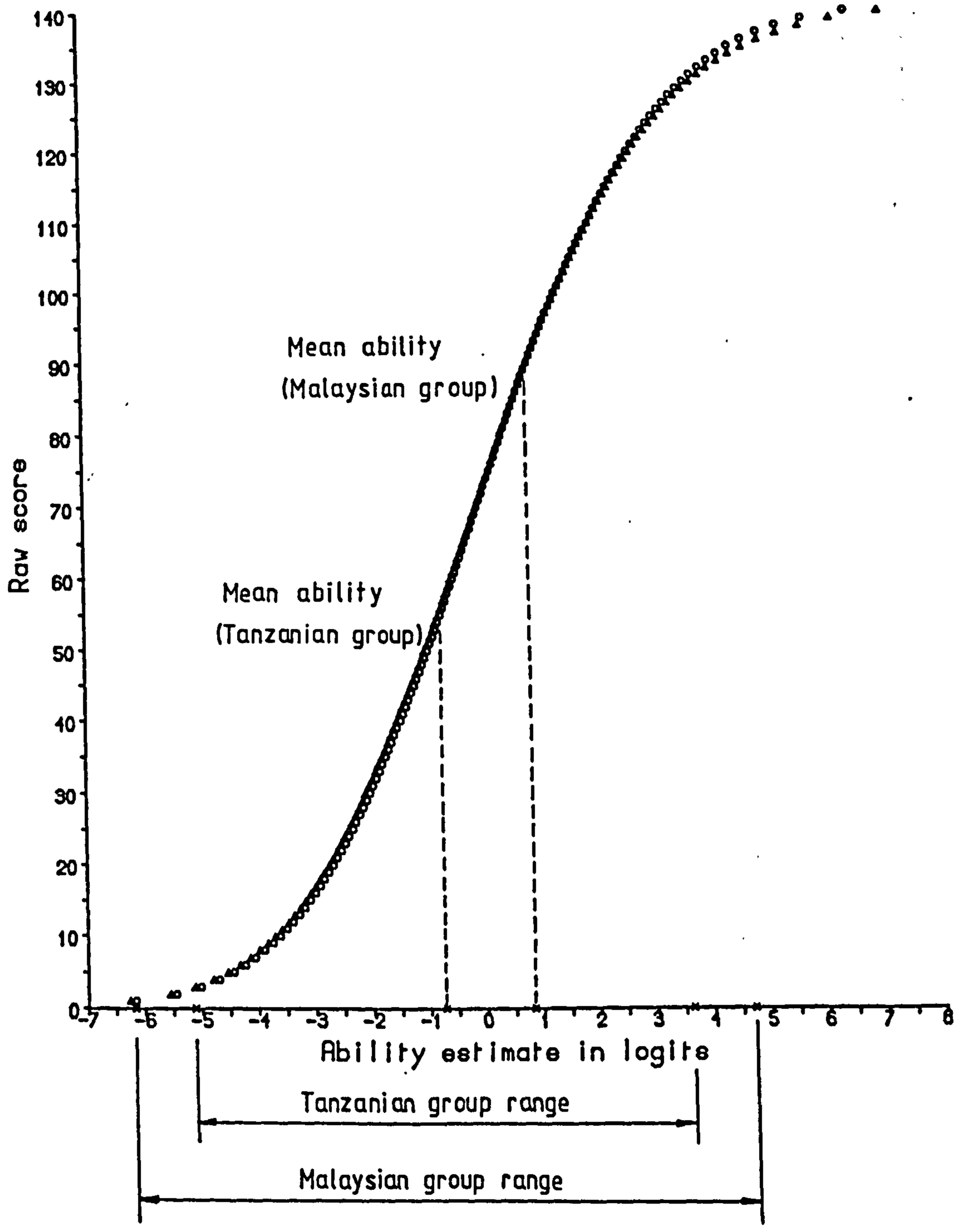


Figure 4.3 Test Characteristic Curves for Cloze-Type Test (Malaysian & Tanzanian Analyses), Showing Group Ability Ranges & Means

Raw Score	Ability Estimate (Malaysian data)	(SE)	Ability Estimate (Tanzanian data)	(SE)
1	-6.17	(1.01)	-6.27	(1.02)
10	-3.62	(0.36)	-3.71	(0.36)
20	-2.67	(0.28)	-2.75	(0.28)
30	-2.01	(0.24)	-2.09	(0.24)
40	-1.47	(0.23)	-1.54	(0.23)
50	-0.99	(0.22)	-1.05	(0.22)
60	-0.54	(0.21)	-0.58	(0.22)
70	-0.09	(0.21)	-0.11	(0.22)
80	0.36	(0.22)	0.36	(0.22)
90	0.84	(0.22)	0.84	(0.23)
100	1.35	(0.23)	1.37	(0.24)
110	1.94	(0.25)	1.96	(0.26)
120	2.65	(0.29)	2.69	(0.29)
130	3.66	(0.36)	3.78	(0.39)
140	6.50	(1.03)	7.09	(1.09)

Table 4.5 Ability Estimates for some Raw Scores, Calculated in Separate Analyses

4.3.2.2 Person Fit

As was indicated in Chapter 2, the analysis of person fit is carried out in order to identify persons whose observed response patterns differ markedly from those predicted by the model, given the difficulty estimate calculated for each item and the ability estimate calculated for each person. It was also noted that the person fit t-statistic referred to here is based on the information-weighted squared standardized residuals, summed for each person across all the items, divided by the sum of information, and converted to a t-test. This statistic thus provides a summary of the discrepancy between each person's observed and expected response pattern.

For this analysis, a maximum acceptable value of 2 for this statistic was set, following the suggestion of Wright, Mead and Bell (1980:15), who describe this as "a good working value" which will "clean most of the implausible response patterns out of the calibrating sample." The value of 2 corresponds to 2 standard deviations above the mean for the theoretical distribution of the t-statistic, which has a mean of 0 and a standard deviation of 1. Where the observed standard deviation is lower than 1, it may be necessary to calculate a new, lower limit for t (see Wright et al., 1980:13). However, the observed mean and standard deviation of the t-statistic for the Malaysian group (-0.15 and 0.99 respectively)

do not differ greatly from the theoretical values, and so the suggested limit of 2 seems suitable here.

The total fit t-statistics calculated for the Malaysian testees are plotted against their ability estimates in Figure 4.4 below.

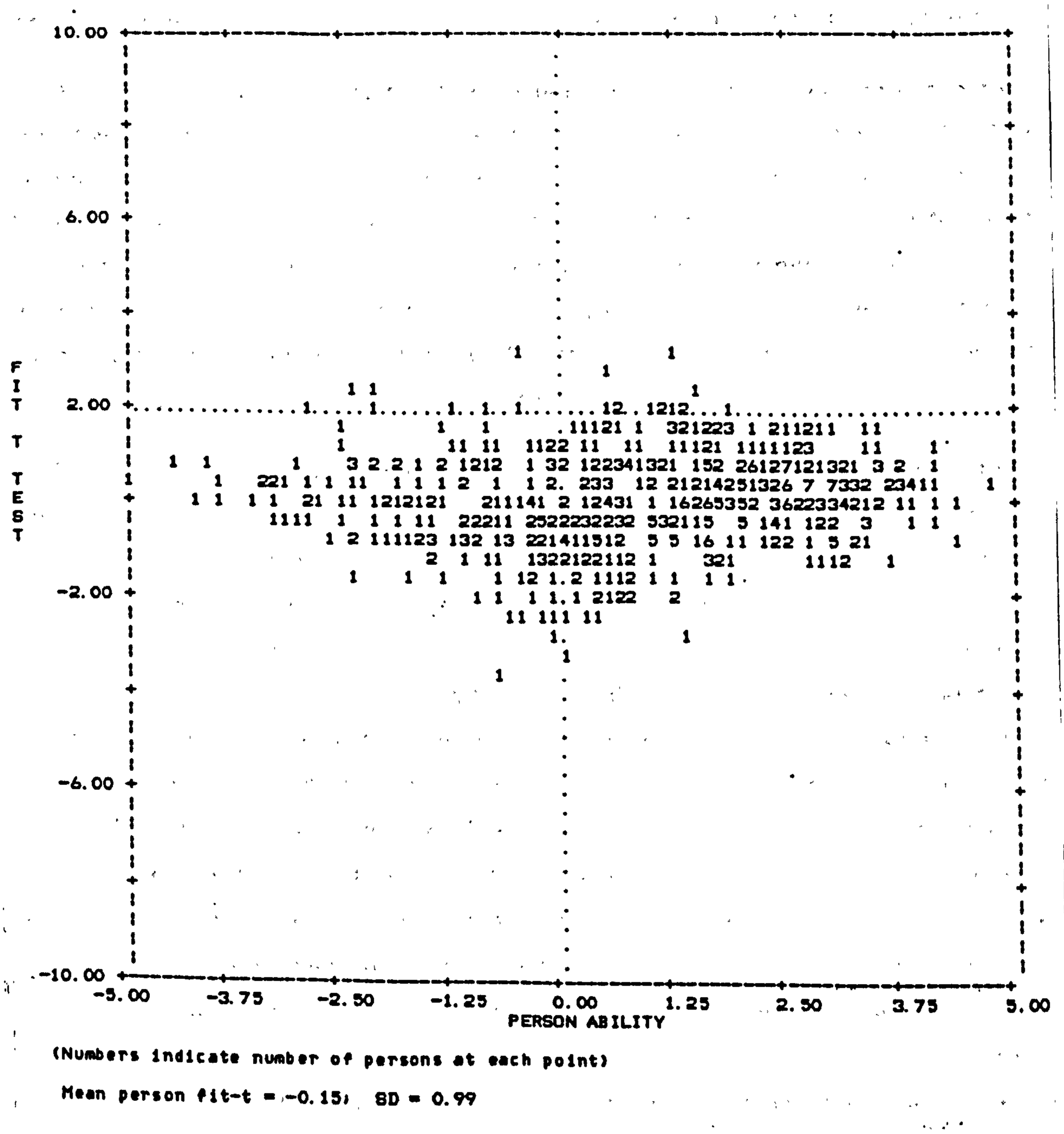


Figure 4.4 Fit t-Test for each Person, Plotted against Ability (Malaysian Testees)

It can be seen from Figure 4.4 that for all but 6 of the 608 persons included in the initial analysis, the total fit t-statistic falls at or below the chosen limit of

2, indicating that the response patterns of most of the Malaysian testees do not show significant departure from model expectation.

If the scores from this test were being used to make important decisions, one might wish to question the plausibility of the response patterns of the 15 persons with fit t-statistics of 2 (i.e. those appearing along the horizontal dotted line in Figure 4.4). However, it is the 6 persons with fit t-statistics of greater than 2 whose response patterns will have caused the most disturbance to the measurement carried out here, and who have therefore been removed from the data set in order for the ability and difficulty estimates to be calculated afresh.

The standardized residuals for these 6 persons on each of the items, together with their ability estimates and fit t-statistics, are set out in Appendix E.5. The persons appear in order of fit, beginning with the most serious case of misfit. Since detailed discussion of each instance of misfit is not the intended focus of this section, an exhaustive analysis of the residuals for each person would not be appropriate here. It is appropriate, however, to provide some illustrative examples of the information that can be obtained from an examination of the residuals and of the responses to which they draw attention.

The pattern of residuals for the person showing the greatest misfit indicates a large proportion of unexpected incorrect answers in the first half of the test (signalled by the accumulation of negative values), and a number of unexpected correct answers in the second half (signalled by the positive values). Indeed, for 41 of the 141 items, the right/wrong score shows at least some degree of departure from expectation, as is indicated by the number of non-zero residuals. The reason for this lack of fit is immediately apparent if one examines the actual answers given by this person: for 46 of the items, including some of the easiest ones, s/he has supplied fillers of 2 or more words, having forgotten, ignored, missed or not understood the instruction that each blank should be filled by one word only. Since the markers were instructed to mark as correct only the (one-word) answers given in the prepared list, these 46 items were automatically marked wrong, even though in some cases the answers formed acceptable sentences. In view of the accumulation of such answers, this person's ability will have been seriously underestimated, thereby making the correct answers given to some of the more difficult items appear surprising when in fact they are not. Although the source of misfit is less obvious for the other persons listed in Appendix E.5, examination of their answers, particularly on items shown by the magnitude of the residuals to have been highly improbable in outcome, can

nevertheless suggest possible explanations in some cases. For example, it would appear from the two most unexpected wrong answers (both with residuals of -5) given by the second person that these resulted from inattention. The items in question were: 'Peter saw (B22) big stick and picked (B23) up', for which this person supplied the words 'some' and 'them'. The consistency of these answers, coupled with the fairly high score (97) obtained overall, and the absence of other errors involving recognition of singular vs plural noun forms, suggests that this person simply failed to notice, or to take account of, the singular noun in this case, rather than that s/he did not have the necessary knowledge of English structure.

Another of the misfitting persons (no. 564) appears to have answered at random, sometimes using words taken from elsewhere in the passage. Without any other evidence, it is not possible to say whether this strategy was adopted because of very low proficiency or because the person did not, or was for some reason not able to, take the test seriously. The fact that s/he answered none of the items in the last two passages could have come about for either reason. For someone whose total score (28) indicates a rather low level of proficiency, this person has made two correct answers identified by the residuals as being highly improbable (items E54 and J112, with residuals of 6 and 9 respectively[2]). It seems unlikely, in a productive test such as this, that these were answered correctly by chance. Alternative explanations would be that some answers were copied from a neighbour, or, if this person is indeed of higher proficiency than the score suggests, that these apparently surprising correct answers are in fact indicative of his/her true level.

The answer pattern for person no. 436 is unusual in that s/he has omitted a total of 58 items, not only at the end of the test (which might have indicated lack of time), but throughout. Indeed, only the fourth passage has been answered completely; between 2 and 8 items have been omitted in each of the remaining 11 passages. Although the incorrect answers given by this person indicate a low level of proficiency (in that they are frequently neither syntactically nor semantically appropriate), it is possible that his/her score has been further, artificially depressed as a result of reluctance to venture answers when not sure.

Other anomalies which come to light via an examination of the residuals concern the marking of the test. For example, it transpires that the apparent failure of person no. 4 on items C32 and G81, for which the residuals are -5 and -4 respectively, resulted from errors on the part of the teacher who scored this

paper. Although the answers given ('neighbour' and 'period') were among those listed on the marking sheet, the scorer marked them wrong. Complete re-marking of this person's answers reveals 5 such errors, only one of which favours the candidate. It is only in the two cases mentioned above that the residuals drew attention to these errors; for the other 3 items the apparent outcomes did not, according to the Rasch model, appear improbable in view of this person's estimated ability and the estimated difficulties of the items. However, the two noteworthy residuals proved sufficient in this case to alert the tester to this source of unreliability.

A further problem which is exemplified particularly clearly by the answers of person no. 4 is that the marking sheet itself has certain deficiencies. In at least 6 cases, this person supplied words which would be perfectly acceptable, but which were not anticipated when the marking sheet was drawn up. These include 'the' for item C37, 'during' for item D39, 'supply' for item H94, 'am' for item J111, and 'bottle' for item L130.

The foregoing discussion serves to demonstrate that the scores (or their Rasch equivalents) for the persons listed in Appendix E.5 cannot be trusted as measures of their English proficiency, and it is therefore appropriate firstly that they should not be used as such, and secondly that the influence of these response patterns on the estimated abilities and difficulties should be removed.

After re-calibration without these misfitting persons, the person fit statistics are calcuated afresh for all of those remaining. In this second analysis of person fit, one person's fit t-statistic has increased to 2.01, i.e. to just above the maximum acceptable value, while all the others still fall at or below this level. One might wish to investigate the nature of the misfit for this person in the manner suggested above; in BICAL, however, no further persons are removed from the data set, and the ability and difficulty estimates yielded by the second analysis are treated as the final ones.

The results of the initial analysis of person fit for the Tanzanian data set are shown in Figure 4.5 below, in which the total fit t-statistic for each person is plotted against his/her ability. The mean and standard deviation of the t-statistic for the Tanzanian group (-0.12 and 1.07 respectively) are again close to the theoretical values for this distribution, and so a limit of 2 is again used.

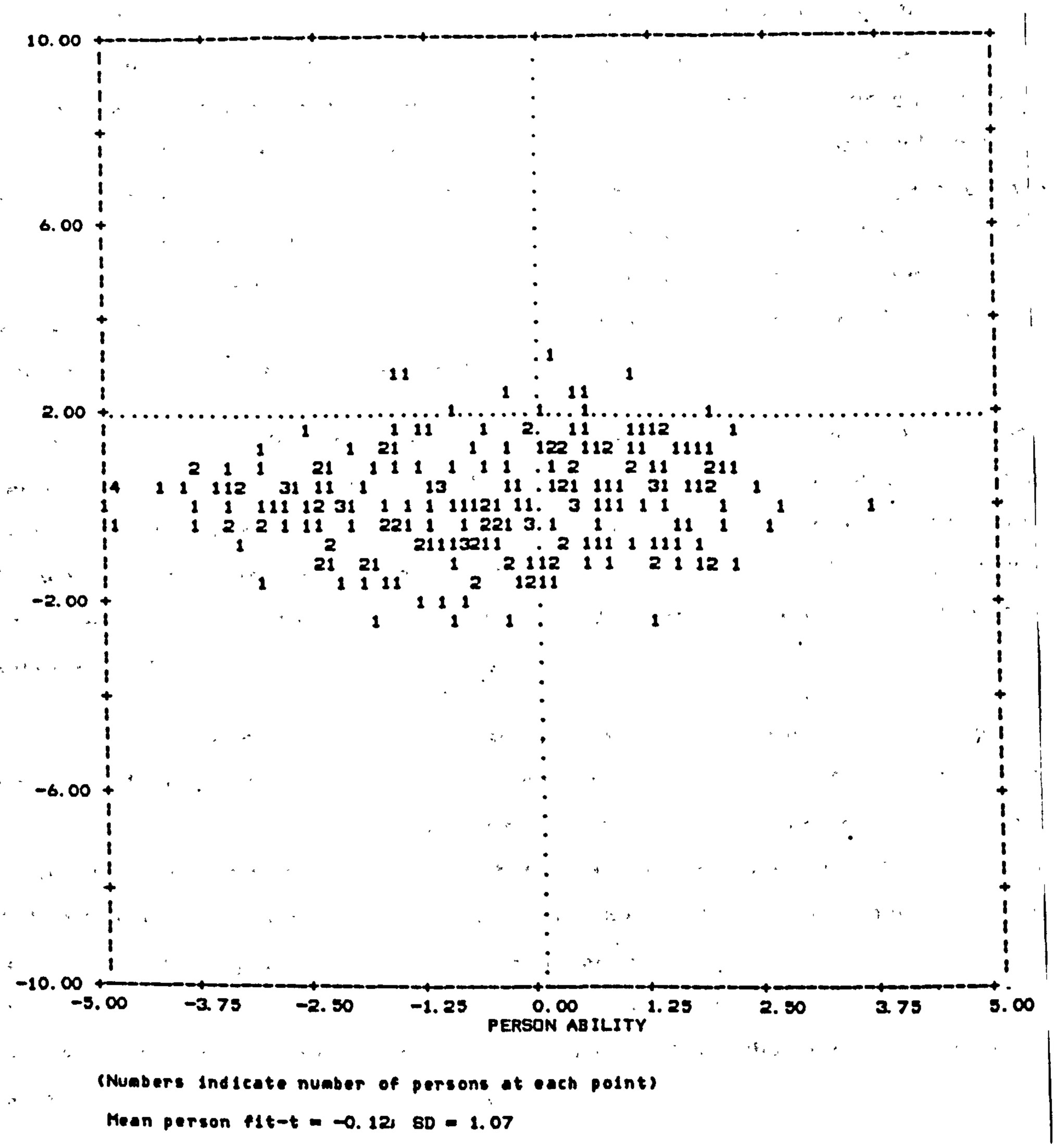


Figure 4.5 Fit t-Test for each Person, Plotted against Ability (Tanzanian Testees)

Figure 4.5 shows that, according to the limit set, 236 of the 243 Tanzanian testees can be considered to have responded largely in accordance with the model's expectations. There are 7 persons for whom the fit t-statistic exceeds 2. As with the Malaysian group, the number of persons misfitting is relatively small; however, as was noted in Section 4.3.2.1, the proportion of misfitters is somewhat greater for the Tanzanian group than for the Malaysian group.

The ability and fit statistics for the 7 misfitting Tanzanian testees, together with their standardized residuals for each item, are set out in Appendix F.5, beginning with the person with the highest fit t-statistic. As was the case for the Malaysian analysis, examination of the residuals and responses for these persons brings to light various inconsistencies, in some cases on the part of the testees and in others on the part of the scorers. For the first person listed (person no. 97), it appears to be a combination of both which has led to the overall impression of misfit: the residuals for certain items in passage F draw attention to consistently erroneous use of present verb forms in this passage when the context clearly demands past forms, a problem to which this person did not seem prone elsewhere in the test; the residual of -3 for item F73, on the other hand, can be attributed to marker error.

The second and third persons listed (nos. 52 and 166) have both omitted items throughout the test (15 and 25 respectively), and hence may not have revealed their true levels. This seems particularly likely in the case of person no. 52, whose total score of 91 suggests that s/he could probably have succeeded on at least some of the omitted items, e.g. A6 and B19, which are of below average difficulty.

Inattention, or perhaps lack of seriousness, might account for the highly unexpected outcome (residual = -8) for person no. 78 on item A1, which is the easiest item in the test. Given that this person answered over half of the items correctly, it does not seem plausible that s/he could not have supplied 'is' in 'It (A1) a big tree' (the actual answer given was 'the'). Although this may simply be a matter of paying insufficient attention on an extremely easy item, one of the answers given for a later item suggests that this person has not taken the test entirely seriously: for the item 'It was barking (B16)' s/he has supplied 'woh! woh!'.

Persons 52 and 64 have both unexpectedly failed on one of the easiest items in the test (item A9), both, it would appear from their answers, as a result of having become confused as to which blank space on the answer sheet corresponded to which item on the question paper. Thus attention has been drawn to a further potential source of inconsistency, arising from the format of the answer sheet: if candidates were required to complete blanks appearing in the passages themselves, rather than on a separate sheet, such errors could largely be avoided, since the answers would not be removed from their context.

Again, then, it can be seen that there is reason to doubt the appropriateness

of these persons' scores or ability estimates as reflections of their proficiency levels, and so their response patterns have been removed from the data set in order for a re-calibration to be carried out. In the case of the Tanzanian group, no further misfitting persons are identified by the analysis of person fit performed after re-calibration.

Thus the person fit statistics, as well as permitting identification of particular persons whose proficiency seems not to have been adequately measured by this test, have also, via examination of the residuals and response patterns, drawn attention to a number of potential sources of unreliability in the test procedure. The particular problems identified in this section suggest that reliability could be improved by extending the marking scheme, by introducing some form of check on the accuracy of scoring, by changing the format of the answer sheet, and by ensuring that testees understand exactly what is required.

4.3.2.3 Item Difficulty Estimates

The Rasch item difficulty estimates calculated for each of the 141 items in the test, using the Malaysian data, are given in Appendix E.2. Two sets of estimates are shown, the first set based on the responses of all 608 persons who scored anything other than zero or full marks, and the second set calculated after the removal of the 6 misfitting persons from the data set. In each case, the difficulty estimates are shown with their associated standard errors.

Comparison of the two sets shows that the removal of the misfitting persons' response data has resulted in only minor changes to the difficulty estimates: the differences between the pairs of estimates for the same item range from 0 to 0.09 logits, with an average magnitude of difference of less than 0.02 logits. Although the differences are negligible, discussion in this section will be based on the second set of estimates, since this is the set which would be used for purposes of item analysis.

The final difficulty scale for this test, represented by the second set of estimates, ranges from −3.86 to 4.79 logits. The mean item difficulty is set to zero as part of the analysis, and the standard deviation is 1.81. The distribution of the item difficulty estimates is shown in Figure 4.6 below.

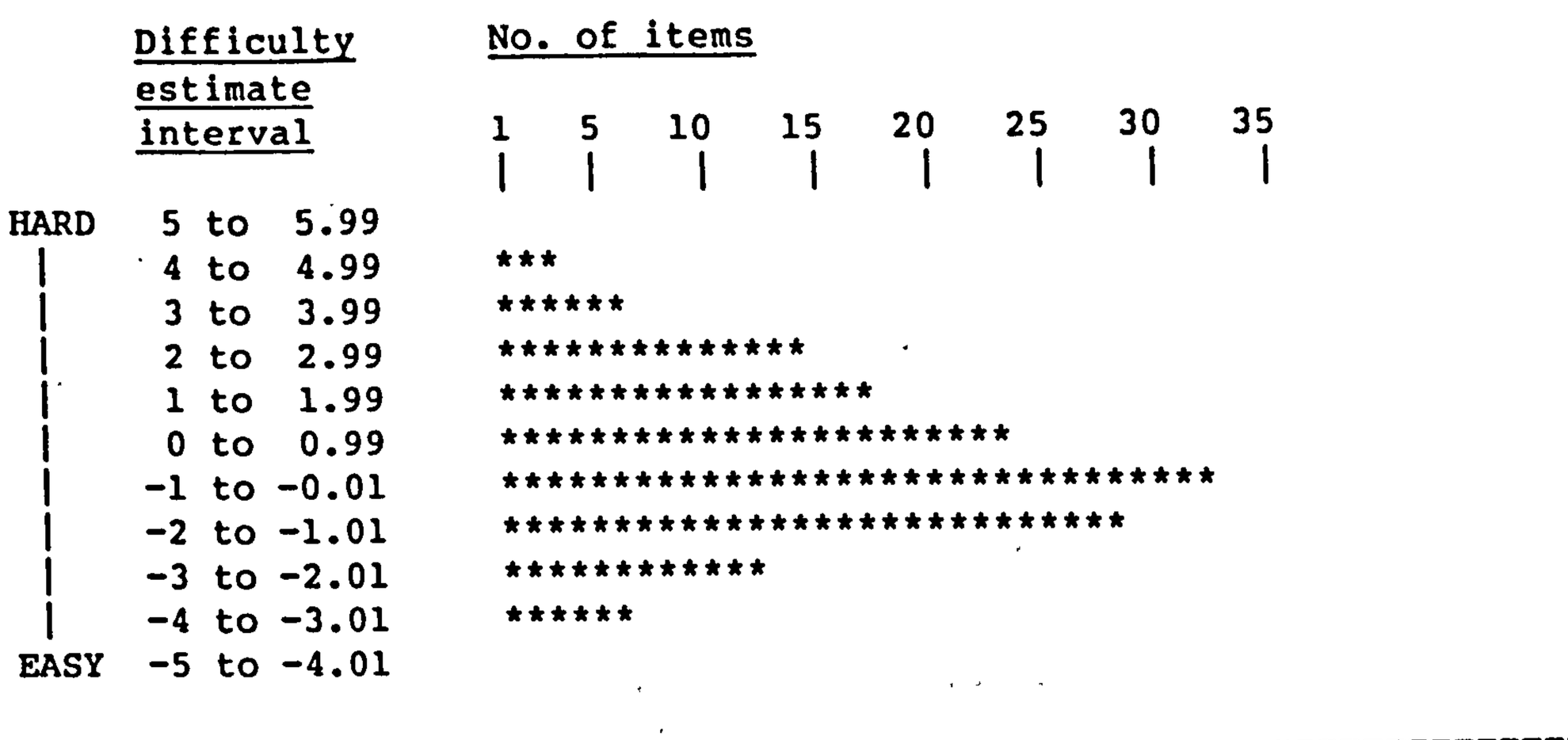


Figure 4.6 Distribution of Item Difficulty Estimates for Cloze-Type Test (Malaysian Data)

The standard errors for these difficulty estimates range from 0.10 to 0.24. They are largest (ranging from .20 to .24) for the 5 easiest items; at the extreme of difficulty they are slightly smaller, increasing only to 0.19. Standard errors are smallest for the 32 items with difficulty estimates between 0.54 logits and 1.87 logits inclusive; again, information is greatest, and hence the standard errors lowest, for the items which are most closely matched in level with the abilities of the persons in the sample, i.e. those answered correctly by approximately 50% of the testees.

The two sets of difficulty estimates obtained from the Rasch analysis of the Tanzanian data are set out in Appendix F.2, the first set calculated using the responses of all 243 measurable persons, and the second set after the removal of the 7 persons identified as misfitting. The effect of 'editing' the data in this way has been greater than for the Malaysian analysis: the differences between the pairs of difficulty estimates in this case range from 0 to 0.42 logits, with an average magnitude of difference of 0.04 logits. However, although more noticeable in the Tanzanian analysis than in the Malaysian one, these differences are again small.

The final difficulty scale yielded by the Tanzanian analysis ranges from -4.16 to 5.94 logits. Again, the mean item difficulty has been set to zero, and the standard deviation in this case is 1.94, which, it will be noted, is larger than that for the Malaysian data. As can be seen from Figure 4.7 below, the range of the

difficulty estimates calculated from the Tanzanian data also extends further in both directions.

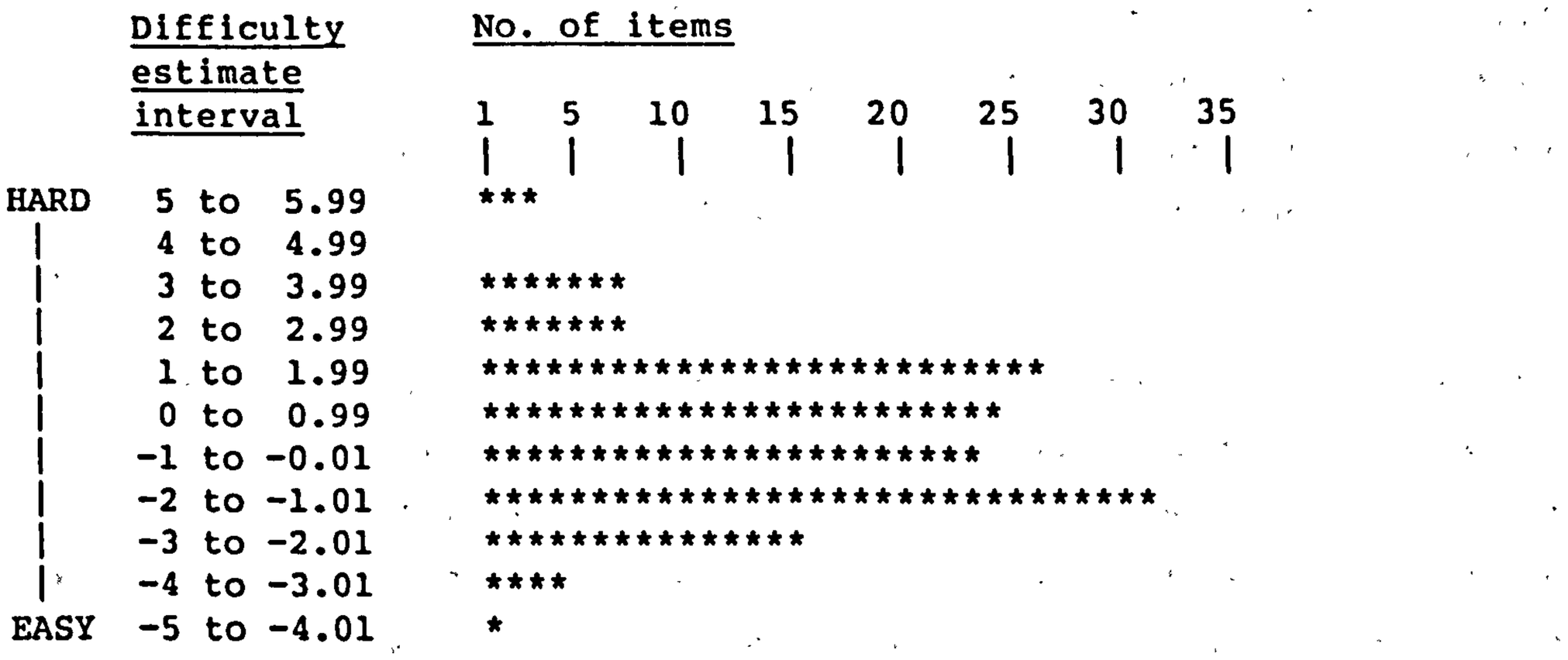


Figure 4.7 Distribution of Item Difficulty Estimates for Cloze-Type Test (Tanzanian Data)

The difference in dispersion of the difficulty estimates obtained from the two testee groups can be seen by comparing Figure 4.7 with Figure 4.6. As regards the particular items falling at the extremes of the scale, the information obtained is the same as that given by the facility values, i.e. that the 10 easiest items for the two groups contain 7 items in common, while the 10 hardest items contain 6 items in common (see Tables 4.7 and 4.8). The product-moment correlation coefficient for the complete sets of item difficulties calculated from the different data sets is 0.87 ($p < 0.001$), again indicating a substantial relationship.

The standard errors for the difficulty estimates obtained from the Tanzanian analysis range from 0.17 to 1.01. The highest values, and hence the least confident difficulty estimates, are associated with the two items at the extreme of difficulty (items I98 and J109, with estimated difficulties of 5.94). At the extreme of easiness, the standard errors are smaller, increasing only to 0.23. The lowest values are those for the 61 items with difficulties from −1.67 to 0.73 logits. The difference between the Malaysian and Tanzanian analyses in terms of the items for which the standard errors are smallest/largest results from the difference between the groups in the proportions of persons passing and failing on each item: the greater the departure from equal proportions passing and failing, the larger the standard error. For the Tanzanian group, the difficulties estimated with the greatest degree of confidence are those for the 61 items with

difficulties from -1.67 to 0.73 logits.

The variation in the magnitude of standard error for different items reflects the fact that items are calibrated more accurately when the person sample is well-matched in ability with the difficulty of the items. Since the Malaysian and Tanzanian samples differ somewhat in ability levels, certain items are better matched to one rather than the other. For the Malaysian group, the items which proved to be most 'on-target' are of generally higher difficulty than those which best suited the Tanzanian group.

To provide an indication of the degree of correspondence between the difficulty estimates for the same item obtained from the separate analyses, some examples of pairs of estimates, together with their standard errors, are set out below in Table 4.6. In order to ensure that these come from throughout the difficulty range, they have been selected by taking the first, and then every 10th item, from the list of items ordered according to difficulty for the Malaysian group.

For the complete set of 141 items, the differences between the pairs of estimates range from 0 to 4.9 logits, with an average magnitude of difference of 0.7 logits. The stability of these estimates between the groups, as compared with that of the facility values, will be considered in Section 4.4.

Item name	Difficulty estimate (Malaysian analysis)	(SE)	Difficulty estimate (Tanzanian analysis)	(SE)
C 25	-3.86	(0.24)	-3.80	(0.24)
D 41	-2.45	(0.16)	-1.79	(0.18)
D 51	-1.84	(0.14)	-1.35	(0.17)
G 83	-1.50	(0.13)	-2.28	(0.19)
H 88	-1.23	(0.13)	-2.01	(0.18)
F 72	-0.79	(0.12)	-1.17	(0.17)
K120	-0.48	(0.11)	0.21	(0.17)
C 36	-0.13	(0.11)	-1.82	(0.18)
L135	0.11	(0.11)	0.43	(0.17)
F 69	0.67	(0.10)	-0.02	(0.17)
F 66	0.88	(0.10)	-0.40	(0.17)
H 94	1.53	(0.10)	1.84	(0.21)
I104	2.12	(0.11)	0.79	(0.18)
H 95	2.81	(0.11)	3.02	(0.29)
J111	4.79	(0.19)	5.24	(0.72)

Table 4.6 Difficulty Estimates for some Items, Calculated in Separate Analyses

4.3.2.4 The Ability/Difficulty Scale

Since the item difficulty estimates and the person ability estimates are all placed on the same scale by this method of analysis, it is possible to view the distributions of person ability and item difficulty together. These are displayed, for the Malaysian data set, in Table 4.7. The column headed 'measure midpoint' sets out the ability/difficulty scale yielded by this analysis, in intervals of 0.2 logits. To the left of this column are the corresponding raw scores, together with frequency counts of the persons falling within the intervals of which the given scale values are the midpoints, and the distances, in terms of standard deviation units, of these ability levels from the mean. To the right of the 'measure midpoint' column are the frequency counts of items whose difficulties fall within the intervals specified, and the names of the items at each level.

By setting out the distributions of persons and items along the variable defined by the test as in Table 4.7, one can gain an immediate impression of the degree to which the persons and items are well-matched in levels. It can be seen that in this case, very few items were close in level to persons with raw scores of more than about 126 (i.e. with ability estimates of greater than approximately 3.2 logits): only 5 items have difficulties greater than 3.2 logits, and

these are separated by a gap of 0.4 logits from the rest of the items. Similarly, the table shows that at the lower end of the scale there are only 6 items which correspond well with the abilities of persons scoring less than about 16. While there were no persons who were too able for all the items in the test, there were, at the opposite end of the scale, 4 persons whose abilities were lower than all of the item difficulties, i.e. for whom there were in effect no suitable items. In general, though, the persons and items in this administration appear to be well-suited, in that for most levels of ability there are a number of items of corresponding difficulty.

PERSON STATS	COUNT	RAW SCORE	MEASURE MIDPOINT	ITEM COUNTS	ITEM NAMES									
	1	136	4.70	1	J111									
+260		135	4.50											
	2	134	4.30	1	I 98									
	5	133	4.10	1	J109									
	1	132	3.90											
	12	130	3.70	2	L133	L139								
	10	128	3.50											
	9	127	3.30											
	19	125	3.10	4	H 89	I101	L136	L138						
	30	122	2.90	2	H 92	H 95								
+180	19	120	2.70	2	C 28	K122								
	30	117	2.50	1	L134									
	23	114	2.30	5	E 59	G 82	J112	J115	K128					
	22	111	2.10	4	I104	J110	K124	L131						
	20	108	1.90	3	G 78	J118	K123							
	25	105	1.70	4	A 4	E 54	G 86	K126						
	31	101	1.50	2	E 64	H 94								
	22	98	1.30	3	F 76	G 79	H 91							
	18	94	1.10	5	A 5	E 62	G 77	I100	I105					
MEAN	21	90	0.90	7	B 16	B 17	B 20	C 27	F 66	K121	K125			
	14	86	0.70	6	D 50	F 68	F 69	I104	K129	L141				
	36	81	0.50	2	A 7	J117								
	25	77	0.30	5	E 61	F 74	H 97	J119	L137					
	18	73	0.10	3	I103	J116	L135							
	24	68	-0.10	10	B 14	C 35	C 36	D 42	D 46	D 49	E 63	F 75	H 90	H 96
	19	64	-0.30	5	C 37	F 70	G 87	K127	L132					
	19	59	-0.50	7	C 29	D 48	E 53	F 67	I108	J113	K120			
	12	55	-0.70	4	D 39	F 65	F 72	I 99	I102	L140				
-180	14	50	-0.90	4	B 15	D 43	I107	L130						
	15	46	-1.10	5	A 6	B 13	B 24	D 44	E 55					
	6	42	-1.30	4	C 34	D 40	D 45	F 71	G 80	H 88				
	13	38	-1.50	4	E 56	E 58	F 73	G 81	G 83	H 93				
	9	34	-1.70	7	B 12	B 19	C 26	C 30	E 57	E 60	G 85			
	7	31	-1.90	4	C 32	D 47	D 51	J114						
	5	27	-2.10	6	B 21	B 22	B 23	D 38	D 52	G 84				
	10	24	-2.30	1	B 18									
	7	21	-2.50	3	A 3	A 11	D 41							
-260	2	19	-2.70	2	A 8	C 31								
	6	16	-2.90											
	4	14	-3.10	1	A 9									
	4	12	-3.30	2	A 2	A 10								
	2	11	-3.50	2	A 1	C 33								
	1	9	-3.70											
	2	8	-3.90	1	C 25									
	1	7	-4.10											
	1	6	-4.30											
-340	1	5	-4.50											
		4	-4.70											
			-4.90											
		3	-5.10											
			-5.30											
		2	-5.50											
			-5.70											
			-5.90											
	1	1	-6.10											
-420			-6.30											

No. of items = 141; No. of persons = 402 (4 misfitting persons omitted)

Table 4.7 Ability/Difficulty Scale from Analysis of Malaysian Data

Table 4.8 below shows the equivalent information obtained from the analysis of the Tanzanian data set.

PERSON STATS	COUNT	RAW SCORE	MEASURE MIDPOINT	ITEM COUNTS	ITEM NAMES						
			5.90	2	I 98	J109					
		138	5.70								
			5.50								
		137	5.30	1	J111						
			5.10								
		136	4.90								
		135	4.70								
+3SD		134	4.50								
		133	4.30								
		132	4.10								
		131	3.90	2	A 4	H 92					
	1	129	3.70	2	K122	L130					
		128	3.50								
		126	3.30								
		124	3.10	3	H 95	L138	L139				
+2SD		122	2.90								
	1	119	2.70	2	I101	L136					
	1	117	2.50	1	L133						
	1	114	2.30	3	B 20	E 64	I105				
	3	111	2.10	1	L134						
	8	108	1.90	8	H 91	H 94	I100	I106	J110	J112	K124	K128
	7	105	1.70	4	E 62	H 97	J118	L131				
	7	101	1.50	6	G 77	G 78	G 79	G 82	J115	K123		
	12	97	1.30	4	D 46	E 54	H 89	K125				
+1SD	6	94	1.10	4	A 7	D 50	J117	K129				
	6	90	0.90	6	B 16	C 28	H 96	K121	K126	L137		
	7	86	0.70	6	D 42	E 61	F 74	G 86	I104	L141		
	10	81	0.50	3	A 5	K127	L135					
	11	77	0.30	6	E 53	F 75	G 87	I107	J119	K120		
	10	73	0.10	3	B 17	I108	J116					
	9	69	-0.10	2	F 69	F 76						
	10	64	-0.30	6	C 27	C 35	F 66	F 71	I102	L140		
	7	60	-0.50	6	B 15	D 44	D 48	F 68	I 99	L132		
MEAN	11	56	-0.70	4	C 34	F 67	I103	J113				
	9	51	-0.90	5	C 29	C 37	E 57	E 63	H 90			
	8	47	-1.10	8	B 14	B 21	C 32	D 39	D 45	D 49	E 58	F 72
	11	43	-1.30	5	A 6	D 47	D 51	F 70	G 84			
	4	39	-1.50	7	B 12	B 24	C 26	D 38	D 52	E 55	G 80	
	6	36	-1.70	4	C 33	D 41	E 60	H 93				
	9	32	-1.90	7	B 19	B 23	C 30	C 36	D 40	G 85	J114	
	5	29	-2.10	3	A 3	E 56	H 88					
	8	25	-2.30	5	A 8	B 22	D 43	F 73	G 83			
-1SD	7	23	-2.50	2	B 13	F 65						
	7	20	-2.70	3	A 2	B 18	E 59					
	6	17	-2.90	2	C 31	G 81						
	4	15	-3.10	1	A 9							
	3	13	-3.30	1	A 10							
	8	11	-3.50	1	A 11							
	1	10	-3.70	1	C 25							
	4	8	-3.90									
	1	7	-4.10	1	A 1							
-2SD	1	6	-4.30									
		5	-4.50									
	5	4	-4.70									
			-4.90									
	1	3	-5.10									
			-5.30									

No. of items = 141; No. of persons = 236 (7 misfitting persons omitted)

Table 4.8 Ability/Difficulty Scale from Analysis of Tanzanian Data

The general impression of the match between persons and items is similar to that for the Malaysian analysis: in general, the persons and items are well-matched in levels, but again they correspond less well at the upper and lower extremes of the ability/difficulty scale. For the Tanzanian group, however, the items for which there are fewest persons of corresponding levels are those at the extreme of difficulty: there are 5 items which proved to be above the level of all members of this group, the 3 hardest of these being separated from the

remaining items by a large gap (approximately 1.2 logits). These items would need to be administered to persons of higher proficiency than those in this group in order for confident estimates of their difficulty to be made.

While the highest scoring persons in the Tanzanian group had rather few suitable items on which to be measured, it is at the lower end of the scale that matching has been least successful. For the 28 persons with estimated abilities of −3 logits or less, there are only 5 items of corresponding levels, and there are 7 persons whose abilities are lower than the level of the easiest item.

Comparison of Tables 4.7 and 4.8 again shows the difference between the Malaysian and Tanzanian groups in terms of their general levels. The mean ability for the Malaysian group (0.83 logits) is higher than the mean item difficulty (set to zero in both analyses), while for the Tanzanian group it is lower (−0.74 logits).

4.3.2.5 Item Fit

The results of the analysis of item fit carried out on the final (i.e. 2nd) set of difficulty estimates are set out for the Malaysian analysis in Appendices E.3 and E.4. Appendix E.3 shows, for each item, (i) the proportion of correct responses made by those in each of 6 even-sized groups formed by subdividing the complete person sample on the basis of raw score, and (ii) the difference between each observed proportion of correct answers and the expected proportion. For the latter, positive values indicate that the observed proportion was larger than expected, and negative values that it was smaller. The raw score ranges, and the mean ability estimates in logits, for each of the 6 ability groups are shown beneath the listed results.

Appendix E.4 lists, for each item, (i) the between-group fit t-statistic, which provides an index of the extent to which success rates for the 6 ability subgroups conform to expectation, given the estimated person abilities and item difficulties calculated using the whole sample, (ii) the weighted total fit t-statistic, which summarises the agreement between observed and expected outcomes across all the individuals in the sample, and (iii) the Rasch model-based discrimination index. (Of these only the first two are considered in this section; reference to the Rasch discrimination index will, however, be made in Section 4.5.) The items are ordered by their total fit t-statistics, from best to least well-fitting.

Although the fit limit of 2 which was used in identifiying person misfit should, strictly speaking, also be applicable in this section, a somewhat higher limit will be used instead, since the observed standard deviation of the total fit t-statistics for this set of items is considerably greater than the theoretical value of 1. Using the values obtained here, the limit calculated from the mean (-0.57) plus 2 standard deviations (2 x 3.5) would be 6.45; however, although item fit t-values exceeding this level will certainly be indicative of item misfit, it would be prudent also to view with suspicion as many as possible of those with lower values than this. Thus in order to take account of the unusually high standard deviation, but at the same time to avoid presenting an unrealistically favourable impression of the extent to which these data conform to model expectation, values exceeding 3 will in this discussion be treated as indicating at least some degree of misfit.

It can be seen from Appendix E.4 that the total fit t-statistics for 123 (i.e. 87%) of the 141 items fall below this level; this set of items can for the most part therefore be viewed as measuring in a consistent way.

For the 18 remaining items, the total fit t-values range from 3.6 (for item G82) to 11.71 (for item B13). Again, an exhaustive analysis of each case of misfit is not appropriate here; however, some illustrative examples of the nature of the item misfit identified, and some suggested explanations of the underlying causes, are presented below.

Item B13 shows the greatest inconsistency across the sample as a whole, and, as is evident from its large between-group fit t-value (15.05), also shows a marked departure from model expectation in terms of the proportions of correct answers elicited within each of the 6 ability subgroups. As is indicated by the figures in the 'Item Characteristic Curve' table in Appendix E.3, the success rate on this item for the 2 lowest-level groups was higher than that for the 3rd and 4th groups, and equal to that for the 5th group. The corresponding figures in the 'Departure from Expected ICC' table show that more correct answers than expected were made by the 2 lowest-level groups (considerably more, in the case of Group 1), and fewer than expected by each of the 4 remaining groups.

The context in which this item occurs, and the most frequent answers given, were discussed in Section 4.2.2.3, in connection with its low traditional discrimination index; it was also mentioned that a commonly-occurring 'incorrect' answer to this item should in fact have been marked correct. Since this was the answer supplied by most of those in the mid- to upper-level groups who appeared to fail on this item, the reason for the observed pattern of misfit is

clear.

The second least well-fitting item overall, F65 (total t = 7.6), also shows a reversal in the ordering of the 6 ability subgroups in terms of the proportions of correct answers, albeit a less marked one than item B13. Again, though, the 2 lowest-level ability groups have performed better than expected, and all of the 4 remaining groups, at least to some degree, less well than expected. The between-group fit t-statistic for this item, though large (8.06), is nevertheless smaller than for 6 other items which, on the basis of their total fit t-values, show better fit overall. This would indicate that while the subgroup proportions for item F65 are closer to expectation than those for these other items, the particular persons answering correctly within the subgroups were frequently not those expected.

Item F65 occurs in the following sentence: 'Mr. Davey was a very old man and he (F65) very curious', and the only answer specified on the marking sheet is 'was'. Given the parallel use of the same verb earlier in the sentence, it is perhaps not surprising that relatively large proportions of the low-scoring testees were able to give the correct answer. The frequent occurrence of the answer 'is', even among those whose overall scores were relatively high, cannot easily be explained, though, particularly in view of the many past forms occurring throughout the passage. The proportions across the 6 groups indicate that this answer was not, however, given by those in the highest level group, 98% of whom answered correctly. Another answer, suggested by a number of relatively high scorers, was 'looked'. Despite being syntactically correct, this answer would not be acceptable, since the interpretation of 'curious' required by the passage as a whole is that of 'inquisitive' rather than that of 'strange'. However, those who chose it may have wished to avoid the repetition of 'was', or perhaps thought this too obvious an answer. The third least well-fitting item in terms of total fit, no. A7, shows greater misfit between groups than item F65; although the success rates for the two highest-level groups were very close to those predicted by the model, the two groups in the middle of the range did substantially less well than expected. The lowest-level group, on the other hand, did considerably better than expected. It is thus across the first 4 ability groups that the inconsistency is observed.

The context for item A7 is as follows: '"I (A60) see some buffalo in the river," Swami called (A7) Lalita.' The incorrect answers most commonly given (e.g. 'for', 'on', 'upon', 'out', 'at') are words which might, in various other contexts, follow

'call'. Those offering such answers appear to have taken the item and its immediate context to be a separate sentence. It is possible that their failure to understand the relationship between the phrase beginning 'Swami called ...' and the preceding piece of direct speech was contributed to by the punctuation marks which separate the two, and the initial capital letter of the name.

Item C36, which has a total fit t-value of 6.35, and a between-group fit t-value of 9.43, was answered correctly by considerably fewer of those in the second highest ability group than predicted. The success rates for the 3 lowest-level groups, on the other hand, were somewhat higher than expected.

Item C36 appears in the following context: '"... I've brought you (C34) fruit. I said to myself, 'I must (C35) Mrs. Chong some of my (C36)' ..."'. Although this is one of the 5 items for which the list of acceptable fillers was left open, this in itself does not appear to have been the cause of the inconsistency; none of the other 4 such items (B16, H91, I104 and L131) has been identified as misfitting. The explanation seems to be that although there was an obvious answer, 'fruit', suggested by the preceding sentence, some of the higher level candidates offered alternatives such as 'plants' and 'share', which, in view of the wider context surrounding the item, were not accepted.

Where such attempts to answer creatively are common, the difficulty of the item will be over-estimated (since an unrealistically large number of candidates will appear to have been unable to answer it correctly), so that the lower-level candidates, who frequently offer the more obvious answers, will appear to have been surprisingly successful.

Of the remaining items identified in this analysis as showing some degree of misfit across the group as a whole, two (nos. C34 and F67) can be seen from Appendix E.4 to have extremely high values for the between-group fit t-statistic (15.07 and 12.7 respectively). In both cases the 3 lowest-level groups performed better than expected, and the 3 highest-level groups less well; as may be inferred from the between-group fit t-values, the discrepancies were considerable in some cases (the extent of these can be seen from the proportional departures from expectation shown in Appendix E.3).

Performance on item C34 has already been discussed (see Section 4.2.2.3), since it proved also to be a poor discriminator according to the traditional index. Item F67 ('His eyes (F66) still good and his ears were (F67) too') would seem to provide a further example of a phenomenon referred to above: that of the item

for which relatively proficient candidates offer alternatives to the obvious answer, and in so doing lose marks. In this case the only filler accepted was 'good'. The alternatives offered included 'fine', 'sensitive', 'sharp', 'effective', 'fit', 'functioning' and 'perfect'. Some of these would, of course, be perfectly acceptable, and should be added to the marking sheet; even where they are not suitable, however, these answers all demonstrate recognition of the intended meaning of the sentence, and show evidence of a wider vocabulary than the item itself demanded. Thus failure to supply the answer 'good' cannot automatically be taken as indicative of low proficiency, and the contribution of this item to a test of proficiency is therefore questionable.

The issue of the adequacy of the marking sheet, which arose in the discussion of person fit, comes to the fore again in the analysis of item fit. In addition to the example given in the previous paragraph, one finds, among the items singled out for investigation, several more for which answers other than those specified on the marking sheet are possible. For item D40 ('They took (D40) food with them.'), for example, unanticipated but acceptable answers include 'their', 'lovely' and other suitable adjectives, and for item D44 ('All felt (D44) hungry.'), additional possibilities include 'rather', 'terribly' and 'ravenously'. For item E54 ('"Someone (E54) cut it in the rock ..."') an interpretation other than that implied by the specified answer ('has') is possible, so that adjectives such as 'strong' would also be suitable.

This consideration of item fit has thus drawn attention to a number of anomalies in the functioning of certain items as measures of the proficiency of the Malaysian testees, and has shown that there is in some cases reason to mistrust the difficulty estimates obtained. An observation which arises from this discussion, apart from the obvious need to extend the marking scheme, is that the use of simplified texts appears to some extent to have contributed to misfit. Where the simplification has resulted in a text which is unnatural in style, banal or unnecessarily repetitive, testees of high proficiency have tended to supply answers which are more sophisticated than required, and in so doing have sometimes artificially depressed their scores.

It should be noted that it is not necessarily only items with large total fit t-statistics which require investigation; extremely low values for this statistic may also be undesirable (depending on the purpose of the test) in that they can be indicative of excessive discrimination. Wright et al. (1980:85) note, for example, that extreme discrimination can be caused by an "interaction between an

idiosyncracy of the item and a secondary characteristic of some of the testees".

It can be seen from the items listed at the beginning of Appendix E.4 that those with the largest negative values for total fit-t have relatively large positive values for between-group fit. This results from the fact that the lower-level groups have performed even less well than expected, and the upper-level groups even better; i.e. discrimination between the upper and lower levels has been extreme. (The departures from the expected item characteristic curves for these items, shown in Appendix E.3, confirm this.)

In a cloze-type test such as this, extreme discrimination of this kind might result from the fact that subsets of between 10 and 15 items relate to the same passage, or indeed from the interdependence of items within the same passage. However, the first 6 items listed in Appendix E.4 all belong to different passages, and among the first 20 items, no passage other than passage L has more than 2 items. Thus there would appear to be little evidence of a 'passage effect', or of interdependence between items, here. Even passage L, with 5 of its 12 items appearing in the first 20, may show such sharp discrimination not for either of these reasons, but as a result of a combination of the difficulty of some of its items and its position as final passage in the test: some of the relatively low-level candidates may have omitted (because of the general difficulty of the passage), or failed to reach, items of even fairly low difficulty appearing at the end of the test. A passage effect of this kind could be investigated by administering the passages in a different order.

Whether or not the inclusion of items showing extreme discrimination is thought desirable depends not only on the nature of its possible causes, but also on the intended function of the test. For a graded test such as this, administered to learners of a very wide range of proficiency, with the intention of separating them into a number of different levels, items which discriminate sharply at various points throughout the range would seem to be the most effective.

Wright et al. (1980:82) point out that the fit statistics under discussion here are necessarily sample-dependent, but add that items shown to fit for one sample frequently also fit for others. In view of these remarks, it is of interest to compare the item fit statistics from the Malaysian analysis with those from the Tanzanian analysis, to ascertain whether the information obtained is similar.

The item fit statistics for the Tanzanian sample are set out in Appendices F.3 and F.4. Appendix F.3 shows, for each item, the proportions of correct answers observed within each of 6 raw-score groups, and the proportional departures of

these from model expectation. Appendix F.4 lists the between-group fit t-statistic and the weighted total fit t-statistic for each item.

The mean total fit t-value for the Tanzanian group is -0.28, and the standard deviation 2.38. A calculated limit (mean + 2SD) for the identification of serious misfit would in this case be 4.48. However, as in the previous application, it would be advisable to treat items with values exceeding, say, 3 as showing signs of misfit.

It can be seen from Appendix F.4 that 130 items (i.e. 92% of the total) have total fit t-values of less than 3, and can, for the purposes of this discussion, be considered to have functioned in a consistent manner. There are thus 11 items which, according to the limit used here, may be viewed as misfitting.

Taking the particular items identified as misfitting in the two separate analyses, one finds that item B13 is shown in both to be the most serious case of misfit. Item A7, too, is identified in both analyses as showing serious misfit, according to the respective limits calculated. One further item, no. D48, has a total fit t-value of greater than 3 in both cases. However, none of the other items pinpointed are common to both analyses.

Table 4.9 below lists, on the left-hand side, the items with total fit t-values exceeding 3 for the Malaysian group, ranked by total fit, beginning with the most misfitting item. The corresponding ranks of these items for the Tanzanian group are also shown, for purposes of comparison. The right-hand part of the table lists the items with total fit t-values exceeding 3 for the Tanzanian group, again ranked by seriousness of misfit, and shown with the corresponding ranks for the Malaysian group.

Misfitting items (Malaysian analysis)	Misfit rank (Mal.)	Misfit rank (Tanz.)	Misfitting items (Tanzanian analysis)	Misfit rank (Tanz.)	Misfit rank (Mal.)
B 13	1	1	B 13	1	1
F 65	2	18	D 48	2	14
A 7	3	5	D 52	3	38
C 36	4	94	A 8	4	48
L141	5	16	A 7	5	3
D 44	6	40	B 24	6	24
F 67	7	133	D 41	7	45
F 73	8	90	F 74	8	84
H 94	9	17	D 39	9	60
E 54	10	79	B 22	10	49
L136	9	32	E 59	11	91
B 17	12	38			
C 34	13	44			
D 48	14	2			
A 6	15	19			
K123	16	20			
D 40	17	127			
G 82	18	110			

Table 4.9 Misfitting Items Identified in Rasch Analyses of Malaysian & Tanzanian Data Sets

It can be seen from Table 4.9 that apart from the 3 common items mentioned above, the rankings of items by fit differ considerably from one analysis to the other. The two sets of total fit t-values, for all 141 items, show only moderate correspondence, as the product-moment correlation coefficient indicates ($r = 0.49$; $p < 0.001$). For the two complete sets of between-group fit t-values, the relationship is somewhat closer ($r = 0.60$), but again it would not be advisable simply to assume that patterns of fit will be the same for different groups.

This comparison clearly demonstrates the need, mentioned by Wright et al. (1980:82), to check for fit in every application as a matter of routine, and underlines the point that analysis of fit is concerned not with whether or not 'items fit the model' but with the extent to which, when administered as a set to a given sample of persons, they measure in a consistent way.

In comparing item fit for the two samples used here, the difference between them in terms of raw score distributions must be borne in mind; the effect of this is seen in the mean ability levels of the 6 even-sized subgroups formed from the Tanzanian sample, which differ quite widely from those of the subgroups

used in the Malaysian analysis. The mean Rasch ability estimates for the two sets of subgroups are as follows:

	Group 1	2	3	4	5	6
Malaysian analysis	-2.12	-0.33	0.56	1.46	2.28	3.25
Tanzanian analysis	-3.52	-2.04	-1.01	-0.14	0.73	1.78

It can be seen that the mean ability level of the second lowest Tanzanian group is little higher than that of the lowest-level Malaysian group, and that the highest-level Tanzanian group corresponds more closely to the third highest Malaysian group than to either of the 2 highest-level ones. The patterns of misfit within the Malaysian and Tanzanian samples for some of the items listed in Table 4.9 are compared below.

On item B13, the apparent reversal in the ordering of the 6 ability subgroups in terms of proportions of correct answers is more extreme for the Tanzanian group than for the Malaysian group; as can be seen from Appendix F.3, the highest-scoring Tanzanian subgroup appeared, on this item, to have been the least successful of the 6. Those in the highest-scoring Malaysian group, on the other hand, mostly gave the answer specified on the marking sheet, and therefore showed a predictably high success rate. However, although the pattern of misfit differs somewhat for the two samples, the nature of the misfit is the same: as was mentioned in Section 4.2.2.3, the apparent inconsistency in the performance of the Tanzanian testees on this item resulted from the fact that many of the highest-scoring persons supplied the indefinite article instead of the definite article.

On the other item identified in both analyses as seriously misfitting, item A7, the highest-scoring of the 6 Tanzanian subgroups again did considerably less well than expected, whereas the highest-scoring Malaysian subgroup answered this item almost completely correctly. Examination of the Tanzanian testees' answer papers, however, reveals that the incorrect answers most commonly given were frequently the same as those observed for the Malaysian group. Further evidence of the failure to understand the relationship between the item ('... ," Swami called (A7) Lalita.') and the preceding direct speech is found in other incorrect answers suggested by some of the Tanzanian testees: these include 'Miss', 'sister', 'her' and 'hallo'.

On items F65 and C36, which are among those showing serious misfit for the Malaysian sample, the pattern of correct answers for the Tanzanian testees

departs little from expectation, particularly in the latter case, where none of the proportional departures exceeds 0.08. For item F65, the proportional departures are negligible for the 2nd, 3rd, 5th and 6th subgroups, but slightly larger for the 1st and 4th subgroups (0.12 and −0.18 respectively). Examination of the answers given by the Tanzanian testees indicates that for item F65, there were again a number of occurrences of 'is' (instead of 'was'), but never among the 2 highest-scoring subgroups, whose only 'incorrect' answers were 'seemed', which should have been included in the marking sheet, and 'looked', which is syntactically correct but semantically inappropriate. Thus it would appear that the high-scoring Tanzanian testees were for some reason less prone than their Malaysian counterparts (i.e. those in the mid-range subgroups) to the error in use of tenses noted here; whether this resulted e.g. from paying greater attention to the other past forms used in the passage, from making better use of the clue provided earlier in the sentence, or from the influence of the mother tongue, it is not possible to say.

In the case of item C36, the Tanzanian group offered few plausible substitutes for the obvious answer, 'fruit(s)'. With the exception of one occurrence each of 'harvest' and 'trees', the incorrect answers given all showed evidence of failure to understand the context. The impression given, then, is that where an answer was suggested by the surrounding context, those of the Tanzanian testees who were able to retrieve this did not in general attempt to find alternatives.

In view of the general difference in proficiency between the two samples, it is interesting to note that, for both of these items, the percentages of correct answers given by each sample were almost the same; indeed, the success rates were in both cases slightly higher for the Tanzanian sample, which is the lower in level of the two.

Items D48, D52 and A8 all showed serious misfit for the Tanzanian group, but not for the Malaysian group. The proportional departures shown in Appendix F.3 indicate that on item D48 ('They ate (D48) big lunch.'), the lowest-scoring third of the Tanzanian testees performed considerably better than expected, while the highest-scoring third performed less well than expected. The only answer accepted as correct was 'a'; however, examination of the answer papers reveals that a frequent choice of answer among the higher-level Tanzanian testees who received no credit for this item was 'their'. This, though not as idiomatic as 'a', seems acceptable in the context; indeed, the indefinite article may have seemed inappropriate to some of the more proficient persons, since the lunch in question

had already been referred to earlier in the passage. For the Malaysian testees, the only noteworthy departure from expectation on this item was for the lowest-scoring subgroup, who performed substantially better than the model predicted. It does not seem surprising, however, that the low-level persons in both samples should have supplied an article for this blank, and hence that quite a large number should have chosen (or hit upon) the indefinite article. It should also be borne in mind that in view of the artificially high failure rate on this item, its difficulty will have been overestimated.

Items D52 and A8 showed good fit for the Malaysian sample, and incorrect answers were in both cases confined largely to the two lowest-level subgroups. For the Tanzanian sample, on the other hand, there was in both cases a reversal in the ordering of the ability subgroups, with the 2nd and 3rd groups both performing better than the 4th group. The only correct answer for item D52 ('Ali felt very happy (D51) pleased with his visit (D52) the zoo.') was 'to'. .The most common answers given by the Tanzanian testees who failed on this item were 'at' and 'in'. It is interesting to note that these were also relatively common among the incorrect answers given by the low-scoring Malaysian testees, some of whom were of the same raw-score level as those in the 4th Tanzanian subgroup. It is not clear, however, why the success rate on this item did not show a consistent increase across the groups for the Tanzanian sample.

Item A8, which occurs in a question and answer sequence ('"What else can (A8) see?" asked Lalita. "I can ..."'), required the answer 'you'. This answer was indeed given by 90% of the Malaysian group and 72% of the Tanzanian group. However, a surprising number of relatively high-scoring Tanzanian testees supplied the answer 'I', seemingly having failed to realise that the question is addressed to the other person mentioned in the passage. Although not all of the sequences of direct speech in this passage are explicitly attributed to their speakers, the repeated question and answer structure seems to have caused little difficulty to the Malaysian learners. It is again not immediately apparent why it should have been less straightforward for certain members of the other sample, though this would appear to be a further manifestation of the problem mentioned earlier in connection with item A7, i.e. that of recognising the relationship between sentences, or parts of sentences, containing direct speech.

As regards the items with the lowest total fit t-values for the two samples, i.e. those showing the 'best', or most extreme, fit, comparison of the particular items listed at the beginning of Appendices E.4 and F.4 shows that these do not

coincide to any great extent. Indeed, in the first 5 items in each list, there is only one common item (J116), and in the first 20 items listed, only 5 in common. This is as one would expect, since large negative values for the total fit t-statistic reflect extreme discrimination, and this, of course, is sample-dependent.

This section has shown, then, that for both samples, most of the items in this test have together formed a consistent measure of some ability. In the examples of misfitting items discussed, it has in some cases, though not always, been possible to trace the inconsistency to certain tendencies on the part of the testees, to particular features of the items themselves, or to inadequacies in the marking scheme. The comparison of item fit for the two samples has also demonstrated that some items show similar patterns, and types, of misfit for both, while others appear to interact with some characteristic of a particular group of testees, and show misfit for one sample but not the other. There is some evidence, for example, that misfit of the kind which might be attributable to over-sophistication on the part of the testees caused less disturbance for the sample which contained fewer high-scoring persons.

4.3.2.6 Person Separation

Characteristic of the Rasch approach is an emphasis on the calibration of test items rather than whole tests, and the measurement of individuals rather than groups. However, it is of use, particularly where the application involves a set of items customarily used as a complete test, to obtain statistical information relating to the whole data set, and it is for this reason that the 'person separability index' and the 'number of person strata' are included here.

The 'person separability index', as it is termed by Wright et al. (1980), or the 'test reliability of person separation', as it is called by Wright and Masters (1982), is in fact the Rasch model-based equivalent of the K-R20 reliability coefficient. These terms are intended to emphasise the dependence of this index on the variance of ability within the person sample in question. For both the Malaysian and the Tanzanian sample, both before and after the removal of misfitting persons from the data sets, this index is 0.98, reflecting the high variance in person abilities in each case.

The 'number of person strata' is intended to reflect the number of statistically distinct levels of ability into which the test separates the testees (see Wright and Masters, 1982:106). As can be seen from the method of calculation shown in

Appendix A.6, it is based on the root mean square of the standard errors of the ability estimates and on the standard deviation of ability, and statistical distinctness for these purposes is defined as separation of the mean abilities of the bands by an (arbitrary) distance of 3 standard errors of measurement. According to this index, the Malaysian testees (both with and without the 6 misfitting persons) are separated by this test into at least 9, and possibly 10, distinct levels. For the Tanzanian sample (both with and without the 7 misfitting persons), the number of person strata identified is 9.

4.4 Comparison of Traditional and Rasch Analyses

In the first part of this section, the information yielded by the analyses reported in Sections 4.2 and 4.3 above is compared, the discussion focussing in turn on the information obtained regarding the performance of (a) the persons, (b) the items and (c) the test as a whole. In the second part of the section, further comparisons of traditional and Rasch indices of item difficulty are presented.

4.4.1 Information Obtained from Traditional and Rasch Analyses of Cloze-Type Test Data

4.4.1.1 Performance of Persons

That the Malaysian and Tanzanian testee samples differ in their ranges and mean levels of proficiency (as measured by this test) can, of course, be seen from either set of results. Furthermore, since the relationship between the raw scores and the Rasch ability estimates for a given test can be set out in the form of a conversion table, as in Appendices E.1 and F.1, or in the form of a graph, as in Figure 4.3, the one can easily be retrieved from the other. There are, however, several important differences between the traditional and Rasch analyses in terms of the information yielded about testee performance. These differences can be attributed to (i) certain properties of the respective measurement scales, (ii) the way in which measurement error is viewed in the two approaches, and (iii) the attention given to the response patterns of individuals.

The most obvious difference between the traditional and Rasch ability measures, at least as far as many language testers are concerned, is that the former are considerably more familiar than the latter. The raw scores, which range from 0 to 136 out of a possible total of 141, therefore appear more immediately interpretable than the Rasch ability estimates, which are expressed in

unfamiliar units, and range in this application from about -5 to +5. Although the use of decimals and negative numbers could be avoided by performing some appropriate linear transformation on the estimates, the problem of the unfamiliarity of the scale itself is one which can be overcome only by continued experience with measures of this type. Given such experience, however, this difference between the traditional and Rasch ability scales would cease to be an issue.

A major advantage of the Rasch ability scale, notwithstanding its unfamiliarity, is illustrated by Tables 4.7 and 4.8 in Section 4.3.2.4, from which it can be seen that use of this scale allows direct comparison of persons and items in terms of their standing on the measured variable. It is thus possible to determine, for a given person, or for a group of persons obtaining the same raw score, which items were the most informative as regards the person's or group's ability level. An implication of this for the cloze-type test is that it might be possible, provided that the variation in item difficulty within each passage was not too great, to identify the passages which result in the most efficient measurement at given points on the scale. Such information could usefully be applied in the development of an adaptive implementation of this test, so that testees, instead of taking the whole test, could, on the basis either of (a) prior knowledge of their approximate level or (b) their performance on a 'starter' passage of median difficulty, be presented with shorter, better-targeted versions of the test, and yet have their ability estimates reported on the same scale. The traditional measurement scale, by contrast, does not permit the explicit comparison or matching of persons and items.

As regards the information obtained about the performance of particular persons, one of the differences between the traditional and Rasch analyses is evident in the treatment of the 3 persons who scored zero. These presented no problem for the traditional analysis, and, since they had the same raw score, they would simply be viewed as being of the same (low) level. As far as the Rasch analysis was concerned, however, these persons had to be excluded from the data set, on the grounds that estimations of their positions on the ability scale would not be possible without their having each made at least one correct answer. Under the Rasch approach, all that can be concluded about these persons is that they were too low in level to be measured by this test; in order to find out by how much they were too low in level, or indeed to see whether they were of the same level, they would have had to be tested using suitable sets of easier items. A further important difference between the two analyses

concerns the measurement error reported. The standard error of measurement calculated in the traditional analysis is assumed to apply to scores throughout the range, so that the true scores of all of the Malaysian and Tanzanian testees are considered likely to lie within ±4 raw score points of the observed score. A problem with this approach becomes apparent when one considers the 5 persons with scores of 3 or less. Since scores cannot decrease below 0, there seems to be no sensible interpretation of the standard error of measurement at these levels; indeed, any attempt to apply this view of measurement error at either end of the raw score scale leads to the (absurd) conclusion that persons gaining scores of zero or full marks will have been measured with greater precision than those at any other point of the scale, since the possible intervals for the true score will be at their narrowest for persons falling at the extremes.

In the Rasch analysis, on the other hand, the position is reversed, so that any persons with zero or perfect scores are treated as not having been measured at all, and those with scores near to the extremes of the possible range are considered to have been less well measured than those nearer the centre. Furthermore, as is clear from Section 4.3.2.1, error of ability estimation is viewed as varying at different parts of the scale, rather than as being (approximately) constant for all persons. Thus in the results of the Rasch analysis, a separate standard error was specified for each ability estimate, thereby indicating the degree of precision with which the ability of persons gaining a given raw score had been estimated.

A final important difference between the two methods of analysis, as regards the information yielded about testee performance, concerns the attention given to the patterns of responses made by individuals. In the Rasch analysis, each person's observed right/wrong responses were compared with those predicted by the model (on the basis of the calculated estimates), and the differences reported in the form of standardized residuals. An index of the consistency, or plausibility, of each person's response vector, and hence an indication of whether or not their ability estimates seemed trustworthy, was provided by the person fit statistics discussed in Section 4.3.2.2. On the basis of this information, it was possible to identify a number of testees whose scores did not appear satisfactorily to reflect their likely proficiency levels, and who could therefore be seen as not having been properly measured by this test. In the traditional analysis, on the other hand, no attention was given to the responses made at an individual level, there being no mechanism for examining individual person-item interaction, and thus the question of the plausibility, and hence of the appropriateness, of the person

measures obtained was not addressed.

4.4.1.2 Functioning of Items

Again, the initial unfamiliarity of the scale on which the Rasch item difficulty estimates are made may appear to be a disadvantage; difficulty estimates ranging, as in this application, from about -4 to +6 on the logit scale may seem less readily interpretable than the traditional facility values, on their familiar 0 to 1 scale.

For purposes of judging whether the items were in general of the desired difficulty levels for the two complete groups of testees, the results of the traditional and Rasch analyses proved equally useful: in this regard, the distributions of facility values shown in Appendices C.3 and D.3 and in Figures 4.1 and 4.2 provided the same information as the matched Rasch ability and difficulty distributions set out in Tables 4.7 and 4.8. The two different types of analysis also yielded the same information regarding the difficulty orders of the items for the two testee groups; apart from the occasional slight adjustment resulting from the rounding of values, or from the removal of misfitting persons, the ranking of items by their Rasch difficulty estimates is simply the reverse of the ranking by facility value (low values on the two scales having opposite meanings). Thus for purposes of identifying the easiest and most difficult items for the two groups tested here, either index could be used.

However, in order to be able to assess the suitability of items for individual persons or for persons of a particular score level, rather than for the group as a whole, one needs the additional information provided in Tables 4.7 and 4.8; as was indicated in the previous section, the matching of persons and items in this way is not possible under the traditional approach. Furthermore, if, as is usually the case, one wishes to obtain information about the distribution of item difficulties wihout reference to any particular subpopulation, then use of the Rasch difficulty estimates offers an additional advantage. As can be seen by comparing Figures 4.1 and 4.2 with Figures 4.6 and 4.7, the distributions of Rasch difficulty estimates were less affected by the differences between the two person samples than were the facility value distributions; they therefore gave a more consistent indication of the extent to which the intended design of the test had been achieved, and of the areas of the scale in which further items might need to be added.

The correlation coefficients reported in Sections 4.2.2.2 and 4.3.2.3 for the two

sets of facility values and the two sets of Rasch difficulty estimates reflected the same (fairly) close linear relationship in both cases. However, in order to compare the stability between groups of these indices of item difficulty, it is not sufficient to consider only the extent to which the relationship is linear: one needs to know around which line the values are concentrated. Figures 4.8 and 4.9 show, for the facility values and Rasch difficulty estimates respectively, the nature of the relationship between the pairs of values obtained from the two different person samples[3]. The pair of values for each of the 141 items is represented by a cross in these figures, and the degree of stability shown by each pair is reflected in its position in relation to the line marked 'identity line'. The closer the cross to the identity line, the more similar the two values.

In Figure 4.8, almost all of the plotted points fall below the identity line, indicating that the facility values were nearly all lower for the Tanzanian group than for the Malaysian group. As can be seen from the varying distances of these points from the identity line, the differences in some of the pairs of facility values were only minor, while other pairs showed considerable divergence. One point appears on the identity line, indicating that for one item the facility values were (when expressed to two decimal places, at least) exactly the same for the two groups. Only one item (E59), represented by the outlying point in the upper left portion of the figure, had a substantially higher facility value for the Tanzanian group than for the Malaysian group, a result which, it should be noted, is attributable to a serious discrepancy in the marking[4]. The other 4 points falling above the identity line are nevertheless very close to it, indicating only small differences in the pairs of facility values.

Thus although some of the facility values appear to be fairly stable for the two groups, there are many for which the differences would not be negligible if one wished to characterise the difficulty of items without reference to a particular testee group. The average magnitude of difference within the pairs of facility values, which was reported in Section 4.2.2.2, can perhaps best be interpreted with reference to Figure 4.8, since this provides a visual representation of the facility value scale. On average, the differences in the pairs of values correspond to the distance represented by 0.0 to 0.2 on the axes, i.e. a distance which represents 20% of the available scale.

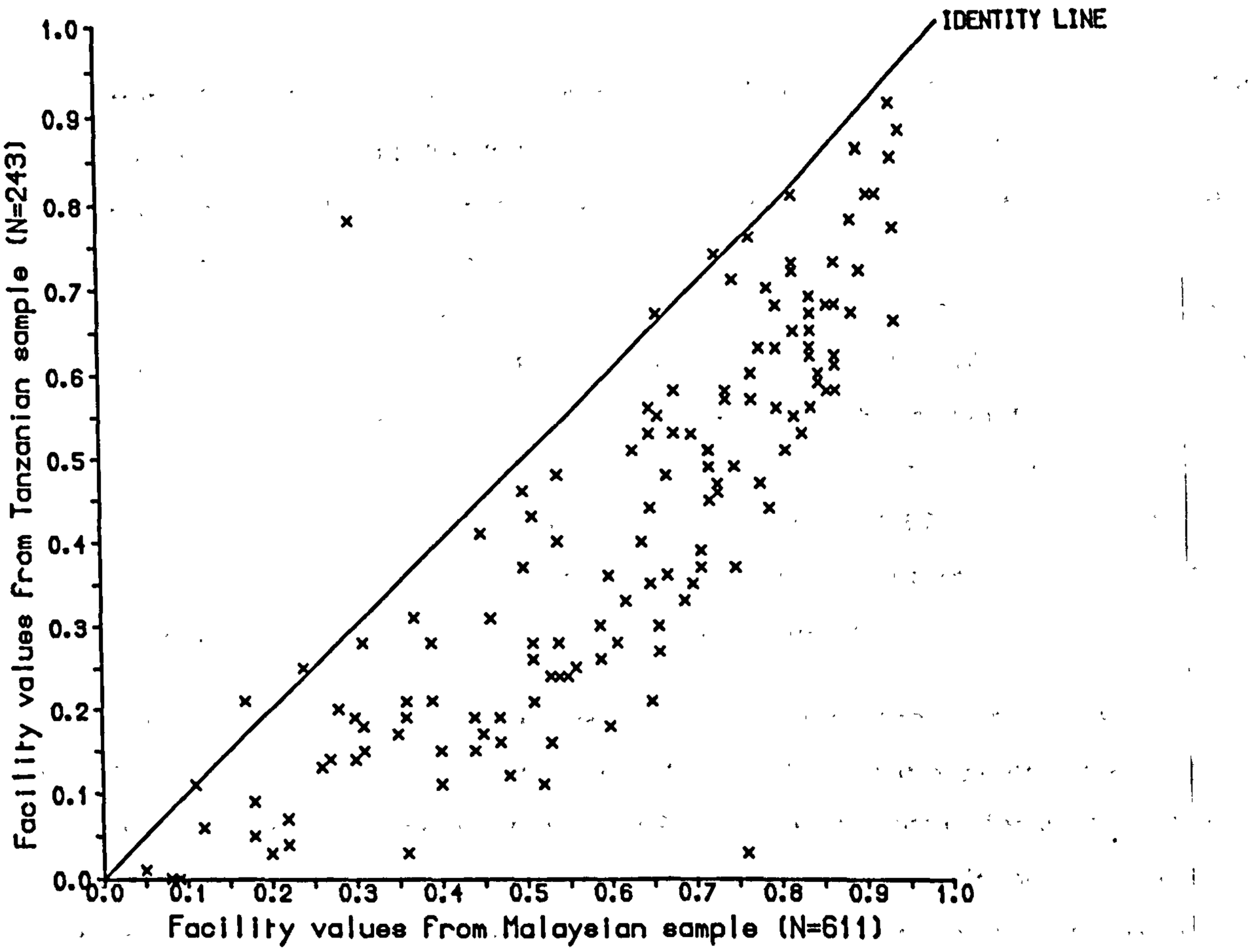


Figure 4.8 Facility Values, Malaysian vs Tanzanian Testees

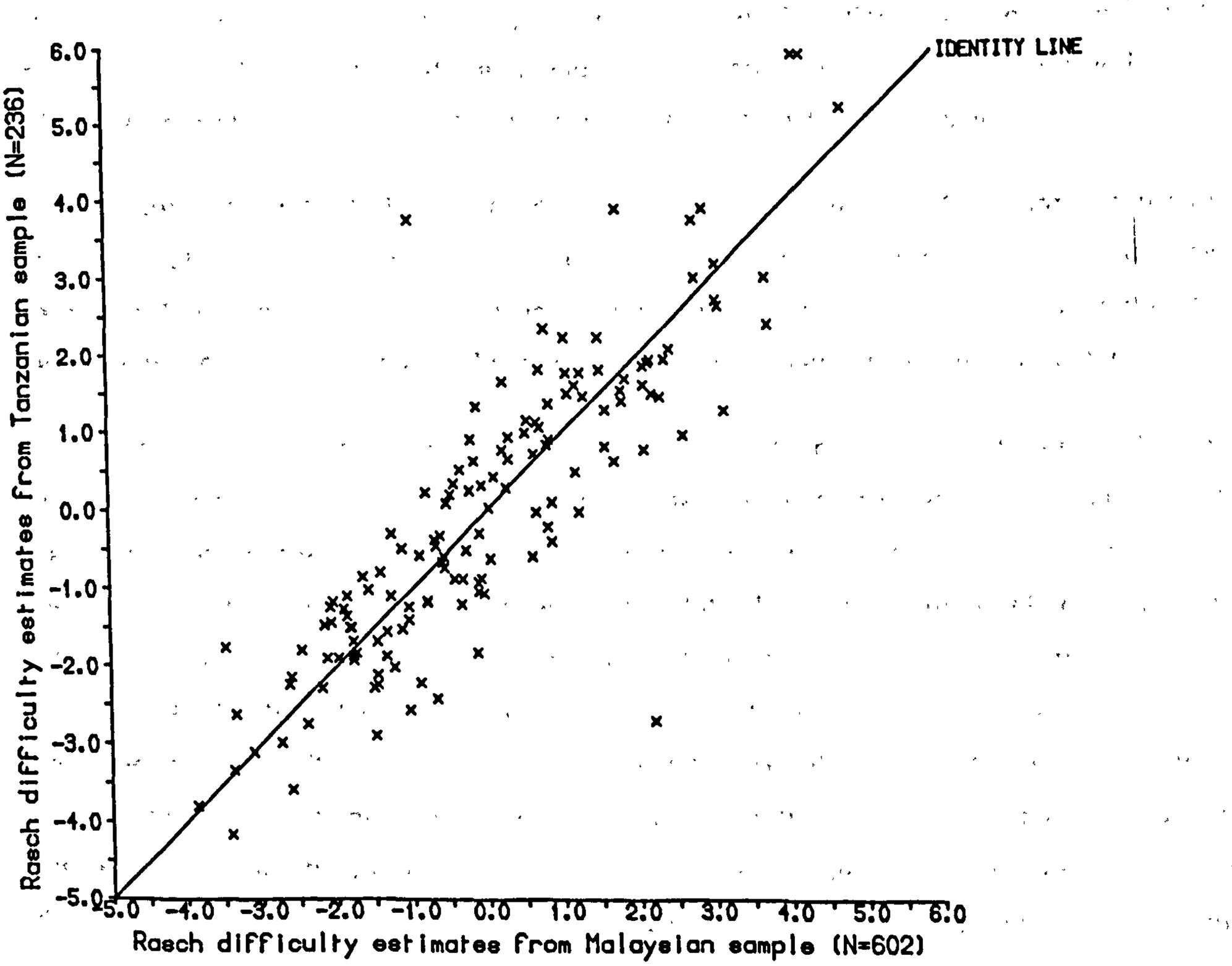


Figure 4.9 Rasch Difficulties, Malaysian vs Tanzanian Testees

In Figure 4.9, which shows the pairs of Rasch difficulty estimates plotted for the two testee groups, roughly equal numbers of points appear above and below the identity line, and the points are in general concentrated quite closely around this line. (For details of the two distant 'outliers', neither of which can be taken seriously, see note 4 at the end of this chapter.) Comparison of Figure 4.9 with Figure 4.8 indicates that unlike the facility values, the difficulty estimates have not been biased by the group ability levels, and that the difficulty estimates are more stable between the two groups. The average magnitude of difference for the pairs of difficulty estimates, reported in Section 4.3.2.3, corresponds to a distance of less than one unit on the scale shown in Figure 4.9, i.e. a distance which represents only about 6% of the length of the axes.

Not all of the points in Figure 4.9 are so close to the identity line that one would be able simply to use either set of difficulty estimates obtained here. It must be remembered, however, that points have been plotted for all of the items in this test, including those for which difficulty estimates could not be made with great confidence because of their extreme easiness or difficulty for one or both of the groups, and those whose difficulty estimates seem unlikely to be trustworthy for reasons of the type suggested in the discussion of item misfit (see Section 4.3.2.5). Nevertheless, Figures 4.8 and 4.9 illustrate that the Rasch difficulty estimates are considerably more stable than the facility values.

It might appear that this greater stability results from the fact that the mean of the Rasch difficulty estimates is set to zero in each analysis, and is thus the same for the two sets of estimates, while the means of the two sets of facility values have been left to vary. This matter will be taken up in Section 4.4.2.

Other differences between the traditional and Rasch indices of item difficulty are analogous with those mentioned in connection with the ability estimates. As was the case for the ability estimates, account was taken in the Rasch analysis of the amount of information available in the data for estimating each of the item difficulties, and a standard error calculated from this was reported for each difficulty estimate. Also, estimation of Rasch item difficulties would not have been possible for items answered all correctly or all incorrectly, just as ability estimation was not possible for the persons who gave no correct answers. Although the facility values listed for the Tanzanian group in Appendix D.2 give the impression that two items (I98 and J109) were answered incorrectly by all members of the group, each of these was in fact answered correctly by one person (facility values to 3 decimal places would be .004) and so estimation of

difficulty, though less confident than for any of the other items, was at least possible.

As far as the assessment of item quality is concerned, it is the indices of discrimination in the traditional analysis and the indices of item fit in the Rasch analysis which fulfil this function. As was indicated in the respective sections, all of these showed at least some degree of sample-dependence, the traditional E_{1-3} index proving to be particularly prone to this. An important difference between the traditional and Rasch indices however, is in their relationship with item difficulty. As was pointed out in Section 4.2.2.3, the items identified by the traditional indices as showing the poorest discrimination were often those which were of extreme easiness or difficulty for the group in question; in order to detect possible inconsistencies in the functioning of items, it was therefore necessary to examine the discrimination statistics in conjunction with the facility values. This relationship between the traditional indices of discrimination and item difficulty is clearly illustrated, for the two data sets analysed here, by the four scatterplots shown below. Figures 4.10 and 4.11 show the E_{1-3} discrimination indices plotted against the facility values for the Malaysian and Tanzanian groups, while Figures 4.12 and 4.13 show the unbiased point biserials plotted against the facility values for each group.

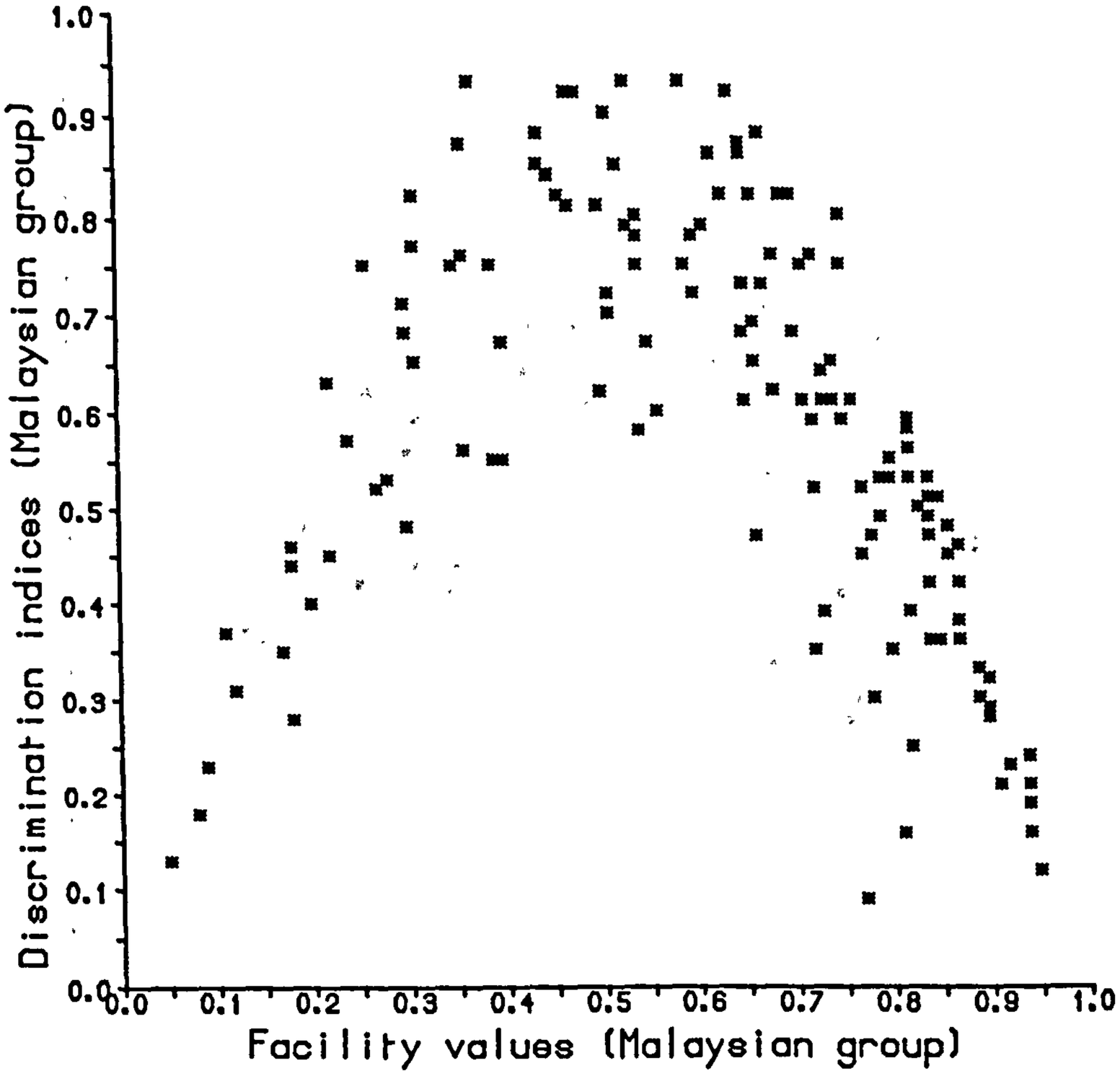


Figure 4.10 Discrimination Indices vs Facility Values (Malaysian Group)

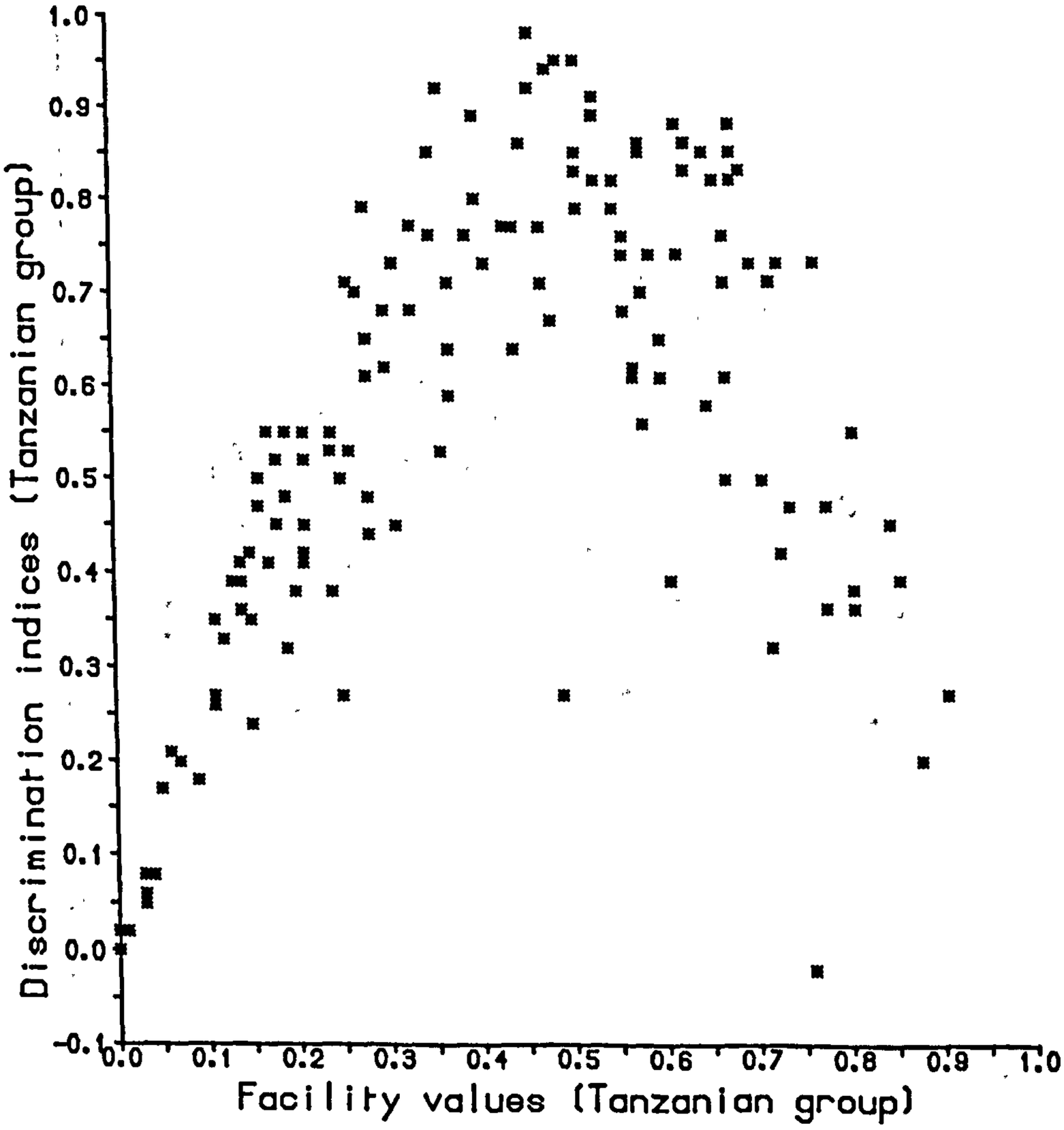


Figure 4.11 Discrimination Indices vs Facility Values (Tanzanian Group)

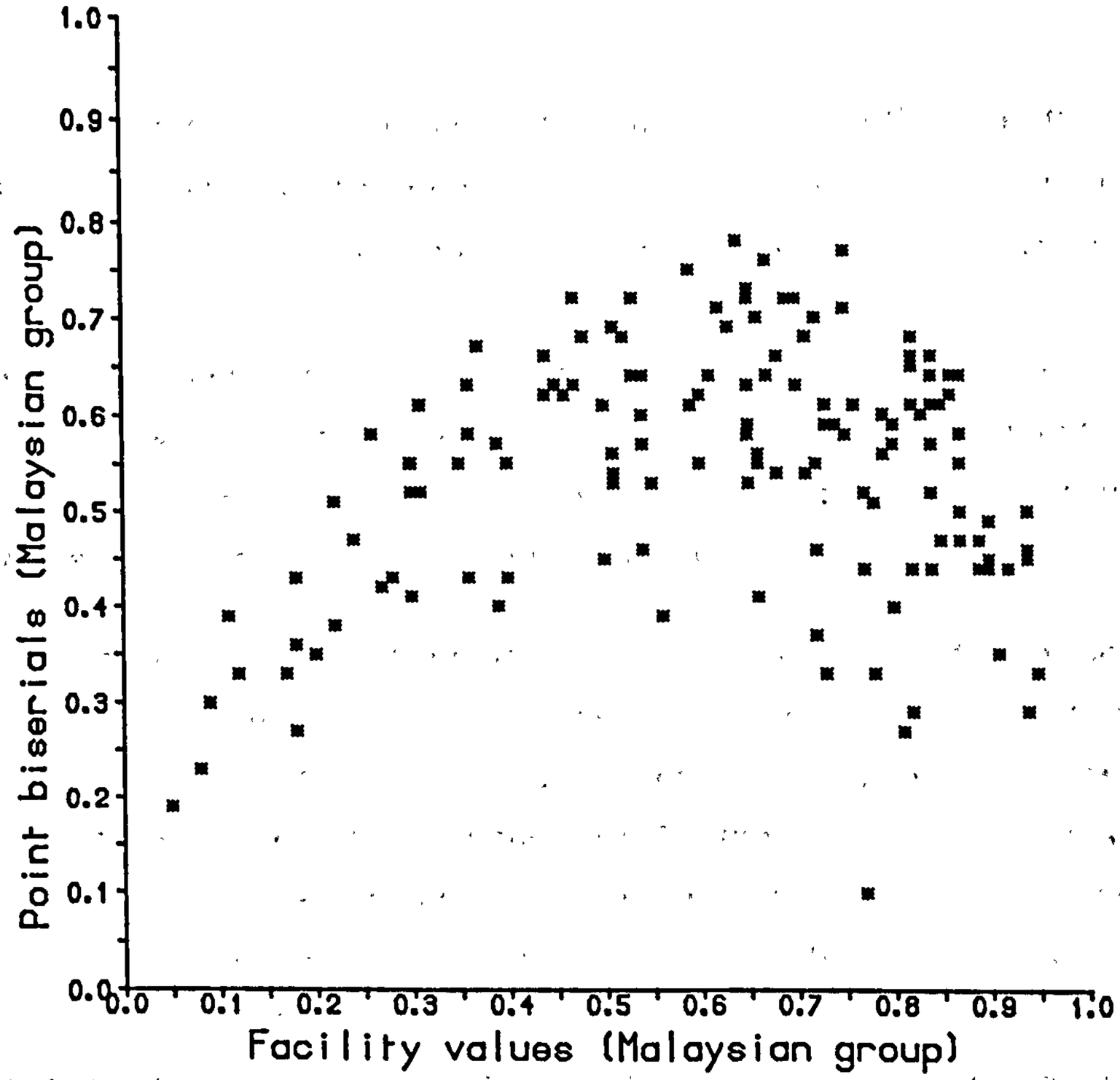


Figure 4.12 Point Biserials vs Facility Values (Malaysian Group)

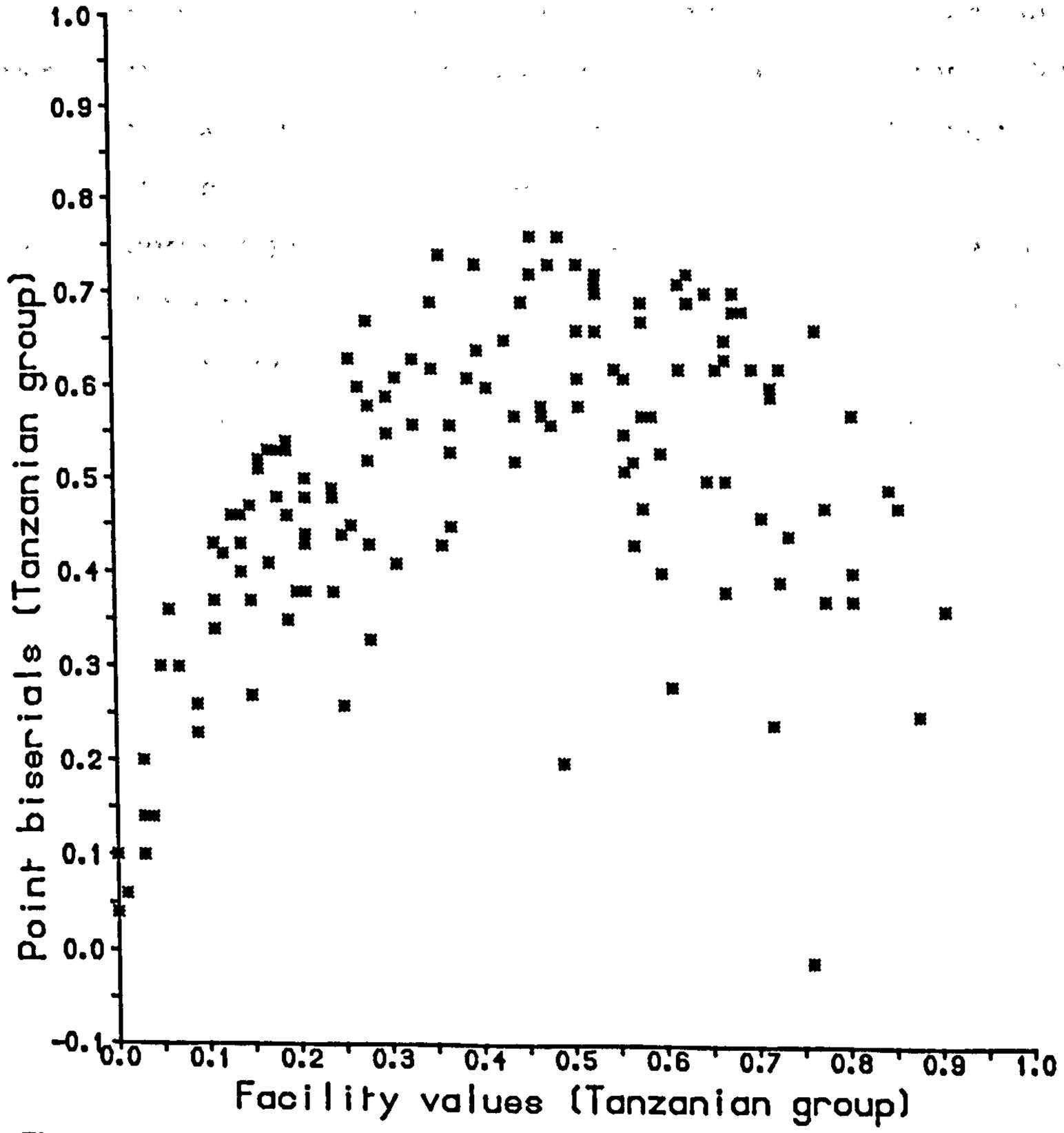


Figure 4.13 Point Biserials vs Facility Values (Tanzanian Group)

It is clear from the inverted v-shaped distributions of the points in Figures 4.10 and 4.11 that discrimination is highest for items close to the mid-point of the facility scale, and that it decreases as items depart from this in either direction. The same tendency is discernible in Figures 4.12 and 4.13, though the narrower range spanned by the point biserials has the effect of making this appear less marked. (The anomalous item B13, it will be noted, is prominent in all 4 figures, since it is placed lower on the vertical axis than any other item, and yet is not at the upper extreme on the horizontal axis.)

The Rasch item fit statistics, on the other hand, are not related to item difficulty in this way. The scatterplots of the total fit t-values and the between-group fit t-values against item difficulty, produced as part of the BICAL output, can be found in Appendix E.6 for the Malaysian group, and in Appendix F.6 for the Tanzanian group. Unlike the four figures discussed above, these scatterplots show no discernible linear relationship between either pair of variables for either group, indicating that extreme easiness or difficulty is not automatically associated with the Rasch indices of item quality.

Given this difference between the traditional discrimination statistics and the item fit statistics, it is to be expected that the 'worst' items identified by the two approaches will not necessarily coincide. Table 4.10 below shows, for each of the two testee groups, the 8 or 9 poorest discriminators identified by the point biserial correlation coefficients, and equal numbers of the most misfitting items identified by the Rasch total fit t-statistics. Items of extreme easiness for the group in question (facility values > 0.9) are marked with a single asterisk, while those of extreme difficulty (facility values < 0.1) are marked with a double asterisk.

Poorest discrimination (Malaysian group)	Most serious misfit (Malaysian group)	Poorest discrimination (Tanzanian group)	Most serious misfit (Tanzanian group)
B 13	B 13	B 13	B 13
J111 **	F 65	I 98 **	D 48
I 98 **	A 7	J111 **	D 5
C 34	C 36	J109 **	A 8
L136	L141	L130 **	A 7
F 73	D 44	H 92 **	B 24
A 1 *	F 67	K122 **	D 41
J109 **	F 73	A 4 **	F 74
		D 48	D 39

Table 4.10 Least Discriminating and Most Misfitting Items (Malaysian & Tanzanian Groups)

Although item B13 is shown on both counts, for both groups, to have given rise to the greatest inconsistency, there is otherwise little similarity between the two lists for each group; indeed, there is in each case only one additional item identified by both analyses. As can be seen from the asterisks, 4 of the 8 items with the lowest point biserials for the Malaysian group fell at one or other extreme of the difficulty range, and 7 of the 9 poorest discriminators listed for the Tanzanian testees were of extreme difficulty for that group. None of the items listed as misfitting was of extreme easiness or difficulty for the group in question, however: as was demonstrated in Section 4.3.2.5, these items were identified only on the grounds of having shown inconsistency of some form in relation to the majority of items in the set.

Since the least adequate items identified by the two approaches show little correspondence, for the reasons given above, it is of interest to see how many of the items identified as misfitting in the Rasch analyses would have given grounds for suspicion using one of the traditional indices. For the 18 items with total fit t-values exceeding 3 in the Malaysian analysis (see Table 4.9 for a list of these), the corresponding point biserials range from 0.1 to 0.46, and appear as follows within the intervals shown:

.1	to	.19	1 item
.2	to	.29	3 items
.3	to	.39	4 items
.4	to	.49	10 items

For the 11 items with total fit t-values exceeding 3 in the Tanzanian analysis, the point biserials are as follows:

-.1	to	-.01	1 item
.2	to	.29	4 items
.3	to	.39	4 items
.4	to	.49	2 items

It can be seen from this that items showing overall misfit tend to have low, or fairly low, point biserials; given that both indices essentially reflect consistency of measurement, this is as one would expect. Whether or not items with point biserials of .4 and above would be regarded as requiring attention, however, is another matter; in view of the 'rules of thumb' sometimes given in testing handbooks, it seems likely that frequently they would not.

A further point arising from the comparison of the traditional and Rasch approaches to assessing the consistency of items concerns the ease with which it is possible to pinpoint the area(s) of the ability range where inconsistency is observed. Although the division of data upon which the traditional E_{1-3} discrimination index is based allowed at least some comparison of the performance of subgroups of different ability levels, the division into 6 subgroups which formed the basis for the analysis of between-group fit was more informative, since (a) it included all the persons in the group, instead of only just over half, and (b) it allowed comparisons of performance throughout the ability range, instead of only at the upper and lower extremes. A further advantage of the Rasch approach in this regard was that, unlike the traditional approach, it allowed the explicit comparison of observed and predicted performance, thereby making available (in the 'Departure from Expected ICC' tables) useful information concerning the extent to which the performance of each subgroup was unexpected or inconsistent on a given item.

A final difference which should be noted is that in traditional item analysis, the view has been that the higher the level of discrimination, the better the item. Under the Rasch approach, however, it is suggested that items showing extreme discrimination should be checked for the possible influence of extraneous factors.

4.4.1.3 The Test as a Whole

The matching of persons and items on the same ability/difficulty scale, mentioned in the two previous sections as an advantage offered by Rasch analysis but not by the traditional approach, also provides valuable information about the test as a whole, in that it allows identification of the levels at which

persons are measured most/least accurately by this set of items.

As regards the 'whole-test' statistics calculated in the two analyses, it was pointed out in Section 4.3.2.6 that the Rasch-based 'person separability index' or 'test reliability of person separation' and the traditional K-R20 coefficient of internal consistency reliability provide the same information. However, there is, perhaps, a subtle difference between them in the emphasis and interpretation implied by their names. As was suggested in Chapter 2, there has been a tendency for the K-R20 to be quoted as though it reflected the quality of the test, irrespective of the persons to whom it might be administered. The names used for the Rasch-based index, on the other hand, emphasise the dependence of reliability on the separation of persons, and thus make the proper interpretation of these indices more explicit.

4.4.2 Further Comparisons of Traditional and Rasch Indices of Item Difficulty

In this section, further attention is given to the comparison of traditional and Rasch indices of item difficulty, since, as was mentioned in Section 4.4.1.2, it may appear that the greater stability of the Rasch estimates resulted simply from the fact that the mean item difficulty was set to the same value for both testee groups. These additional investigations are carried out first using the Malaysian and Tanzanian data sets as before, and then, in order to provide a more stringent check on the stability of the different indices considered, using subgroups of high and low scorers drawn from the larger data set.

4.4.2.1 Indices of Item Difficulty for Malaysian vs Tanzanian Groups

Using the facility values obtained from the Malaysian and Tanzanian data sets, two further traditional indices of item difficulty were calculated for each group. These were:

(i) The facility values expressed in standard deviation units from their respective group means, so that both sets were transformed to a new scale with a mean of 0 and a standard deviation of 1. These new values will be referred to here as 'item z-scores', by analogy with the conventional name for person scores transformed in the same way. The direction of this new scale is, of course, the same as that of the facility values, so that low values are still associated with hard items, and high values with easy ones.

(ii) The facility values treated as proportions representing areas under the normal distribution curve, following the method described by Guilford and Fruchter (1978:458-460). The new values were found from tables of standard

scores corresponding to divisions of the area under the normal curve into a larger and a smaller proportion, and will be referred to here as 'item z-scale values'. Using this method the direction of the scale is reversed, so that low values are associated with easy items and high values with hard ones.

These two additional sets of values can be found in Appendix C.4 for the Malaysian group, and in Appendix D.4 for the Tanzanian group. For both indices, the values obtained from the Tanzanian data were plotted against those obtained from the Malaysian data, in exactly the same way as for the facility values and Rasch difficulty estimates (see Figures 4.8 and 4.9). Figures 4.14 and 4.15 below show the plots of the item z- scores and the item z-scale values respectively.

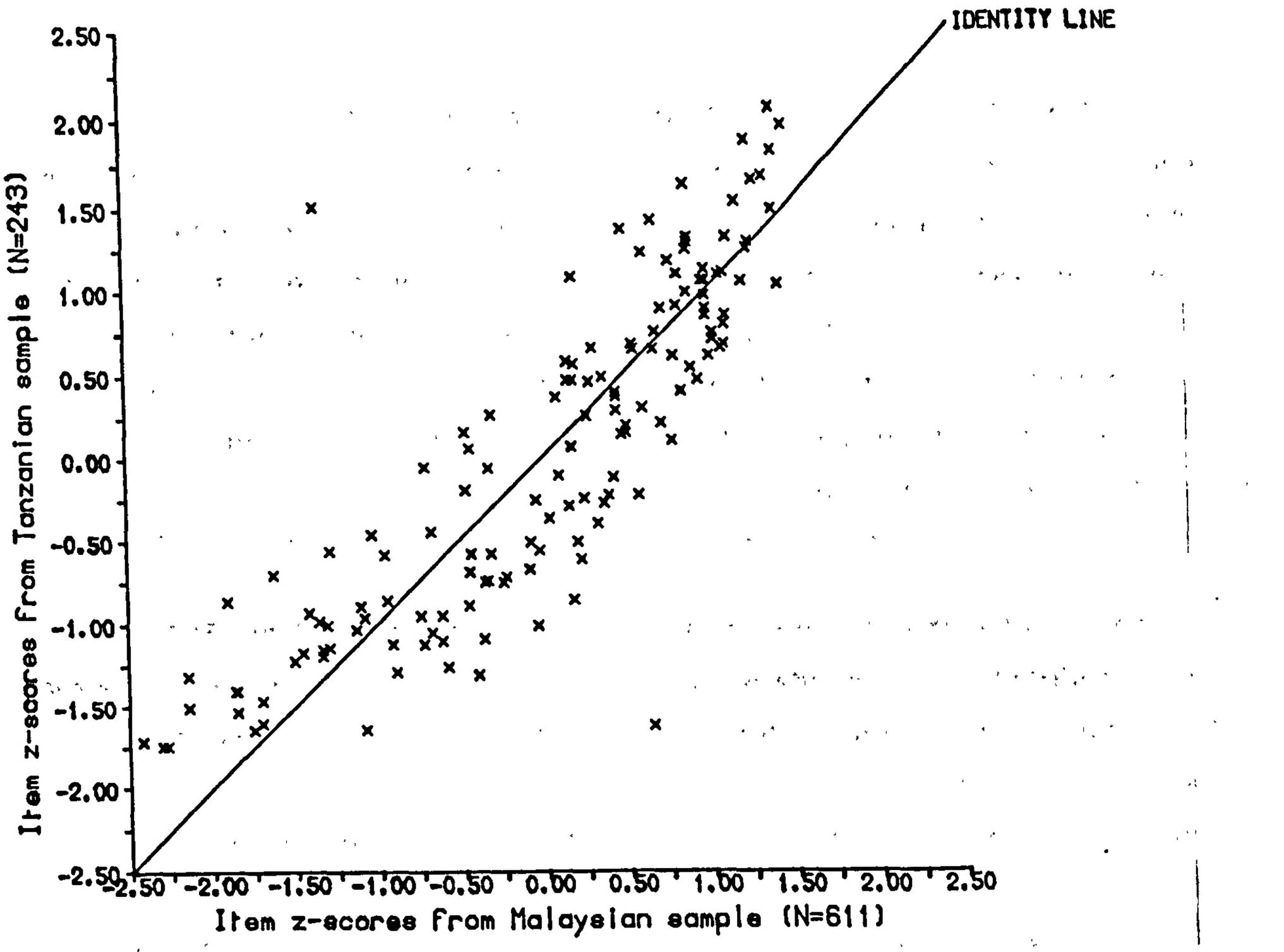


Figure 4.14 Item Z-Scores, Malaysian vs Tanzanian Testees

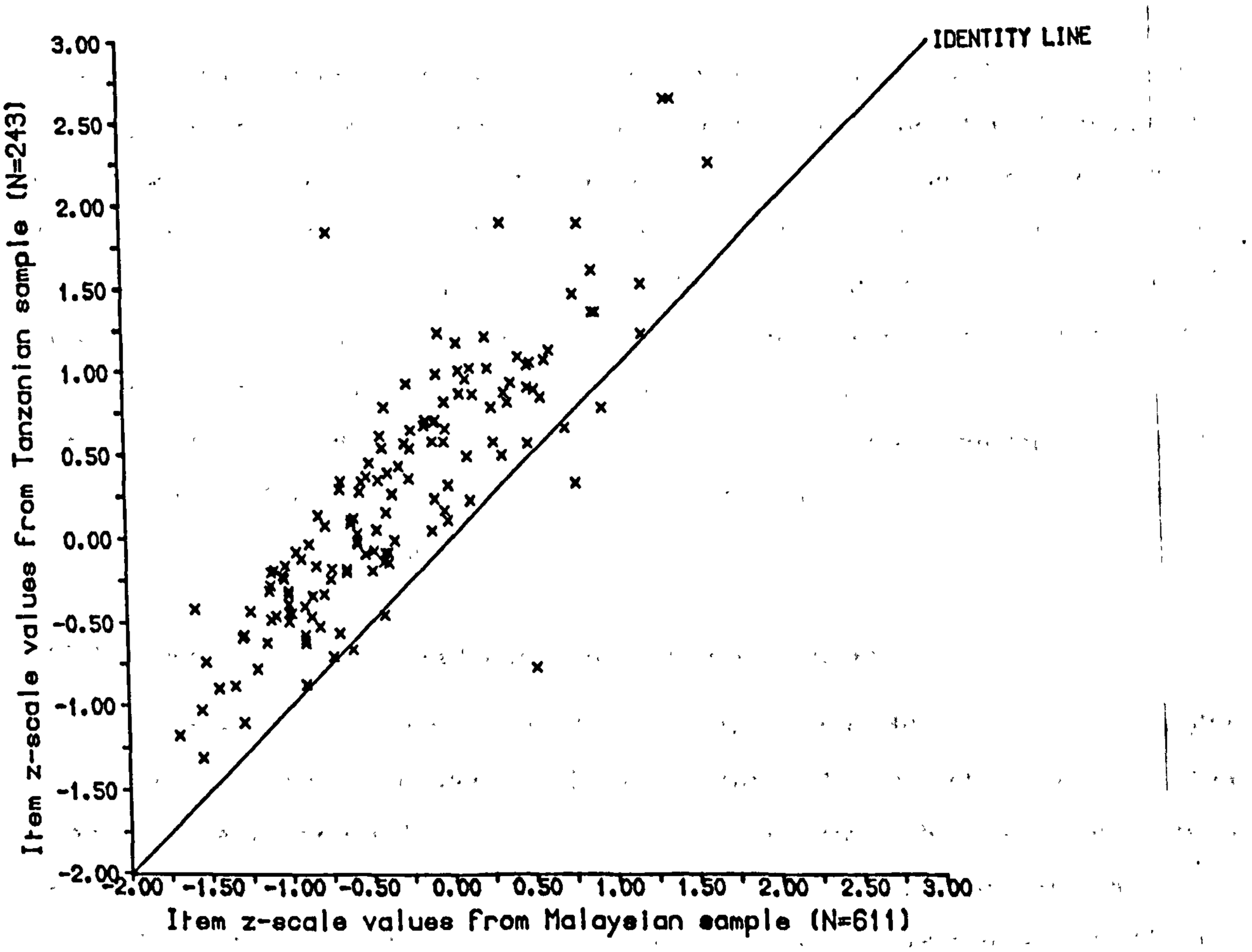


Figure 4.15 Item Z-Scale Values, Malaysian vs Tanzanian Testees

Figure 4.14 shows that the effect of transforming both sets of facility values to a scale with the same mean and standard deviation has been, as one would expect, to centre the points around the identity line. Thus as far as the stability of values between groups is concerned, the item z-scores represent a considerable improvement on the facility values. The use of this simple linear transformation has not, of course, changed the shape of the distribution, as can be seen by comparing Figure 4.14 with Figure 4.8; the only difference (apart from the obvious change in the scale itself) is that the effect of the difference between the mean facility values for the two groups has been removed.

Figure 4.15, on the other hand, shows that although the transformation to the z-scale has brought about a slight change in the shape of the distribution (since this is a non-linear transformation), it has not removed the effects of the difference in ability levels between the two groups. Changing the direction of the original facility value scale has served only to move the points to the opposite side of the identity line. As Guilford and Fruchter (1978:460) note, a further transformation would need to be carried out on one set of these values, in order to make their mean and standard deviation coincide with those of the other set, and hence to centre the points in the figure around the identity line.

Thus of the 4 indices of item difficulty shown in Figures 4.8, 4.9, 4.14 and 4.15, only the Rasch difficulty estimates and the item z-scores offered reasonable stability between the two groups for whom sets of values were compared here. Although similar in this respect, it can be seen from Figures 4.9 and 4.14 that these two indices differ in that the Rasch difficulty estimates are not distributed in the same way as the facility values, while the item z-scores retain the original facility value distribution. The effect of this difference will be seen more clearly in the next section, where the various difficulty indices are compared for testee groups which differ more widely in ability than the Malaysian and Tanzanian groups considered in this section.

4.4.2.2 Indices of Item Difficulty for High vs Low Scorers

As a more stringent check on the stability of the different item difficulty indices, and in order to gain a clearer picture of their characteristics, the response data of two subgroups of testees drawn from the Malaysian sample were re-analysed separately. These subgroups contained (i) the 200 lowest-scoring Malaysian testees (mean raw score 46, score range 0 to 74), and (ii) the 200 highest-scoring Malaysian testees (mean raw score 119, score range 108 to 136). As before, the four item difficulty indices were calculated separately

for each group, and the two sets of values plotted in each case. The results are shown in Figures 4.16 to 4.19 below. (The lists of values can be found, for the traditional indices, in Appendices C.5 and C.6 and, for the Rasch index, in Appendix G.1.)

Figure 4.16 provides a clear illustration of the dependence of facility values on the abilities of the groups from which they are obtained. Very few of the points appear near the identity line, indicating that very few of the facility values were similar for the high- and low-scoring groups: in most cases, the facility values obtained from the high-scoring testees were considerably higher than those obtained from the low-scoring testees, as the clustering of points in the upper left portion of the figure shows.

From Figure 4.17 it can be seen that for this more extreme case, where the two groups differ markedly in proficiency, the item z-scores appear in a less favourable light than in the previous example, where the groups differed relatively little. Although the effect of this transformation has again been to move the points to the centre of the figure, the shape of the original facility value distribution is clearly visible, as are the 'floor' and 'ceiling' effects of the restricted range of possible values: for both the facility values and the item z-scores, these have caused the points to form an almost vertical and an almost horizontal line, with the result that the points are not (as they were in the previous example) spread along the identity line.

In Figure 4.18 the item z-scale values are once again shown to alter the original facility value distribution, but would again require further adjustment to remove the effects of group differences in ability. As with the other traditional indices, a restriction in possible range can be seen to operate. In this case, however, the restriction is somewhat artificial in that the values of ±4 represent practical cut-off points corresponding to areas under the normal distribution curve of 1 and 0; the actual z-scale values for items with facility values of 1 or 0 would, of course, be plus or minus infinity.

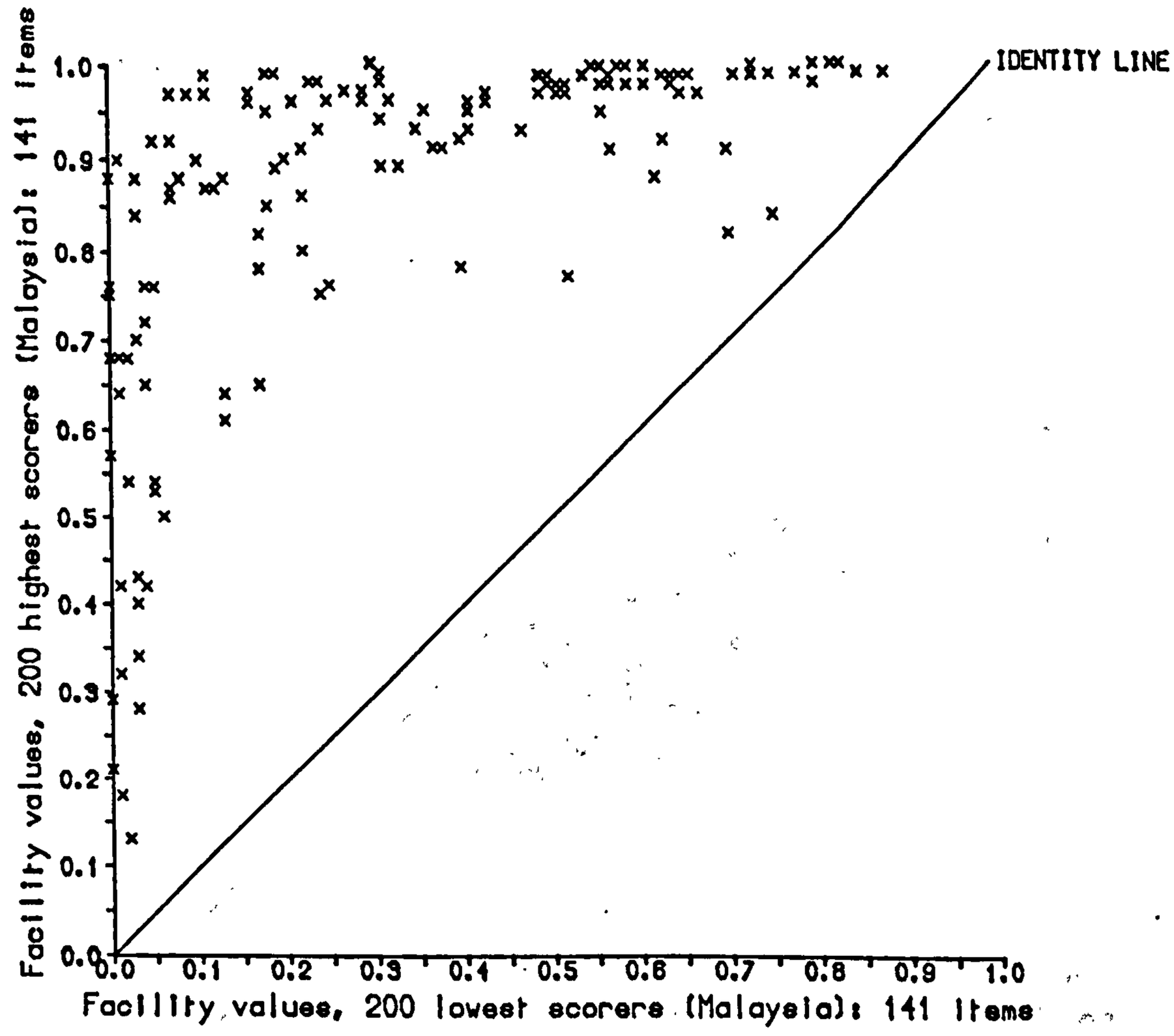


Figure 4.16 Facility Values, High vs Low Scorers (Malaysian Data)

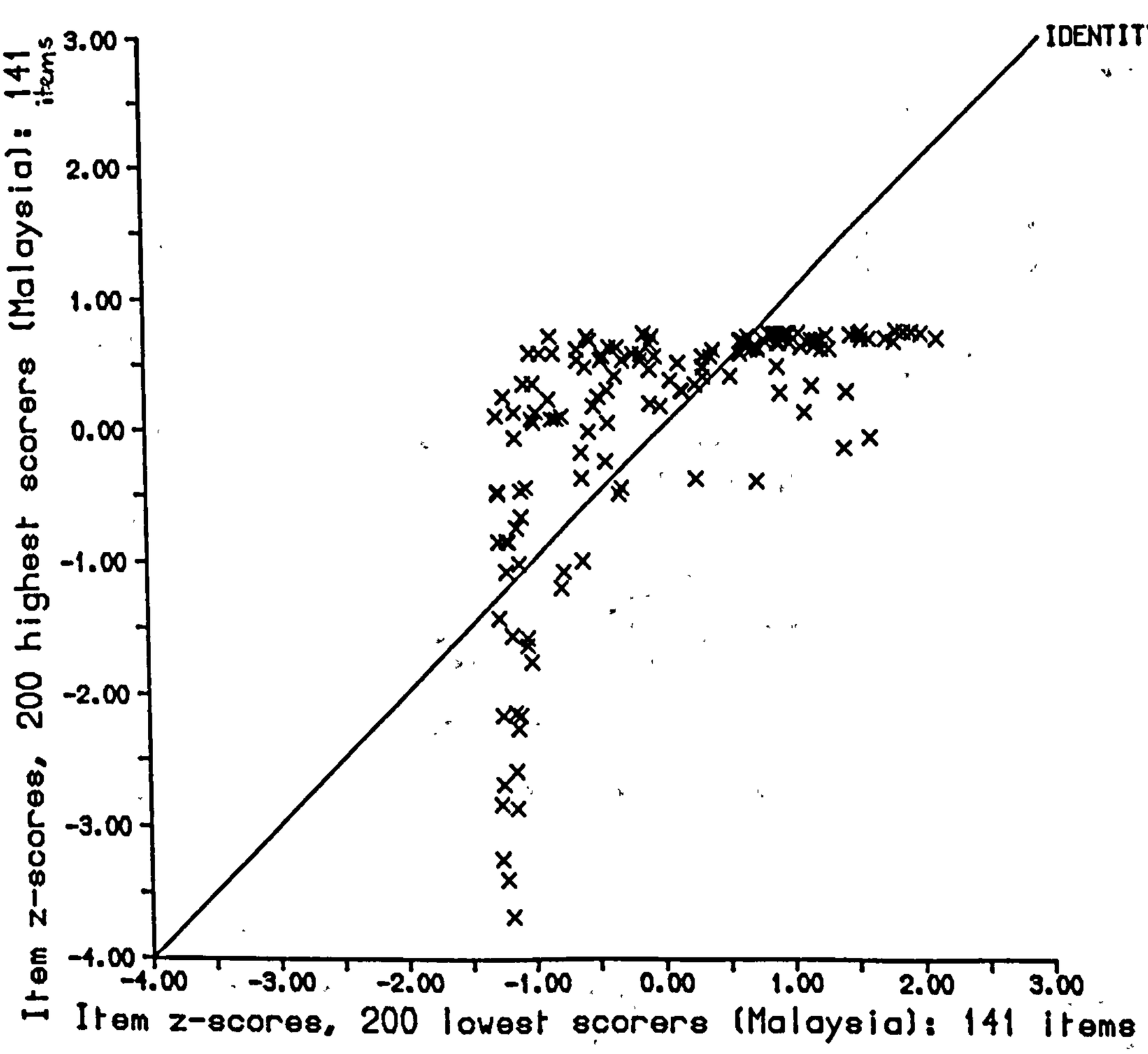


Figure 4.17 Item Z-Scores, High vs Low Scorers (Malaysian Data)

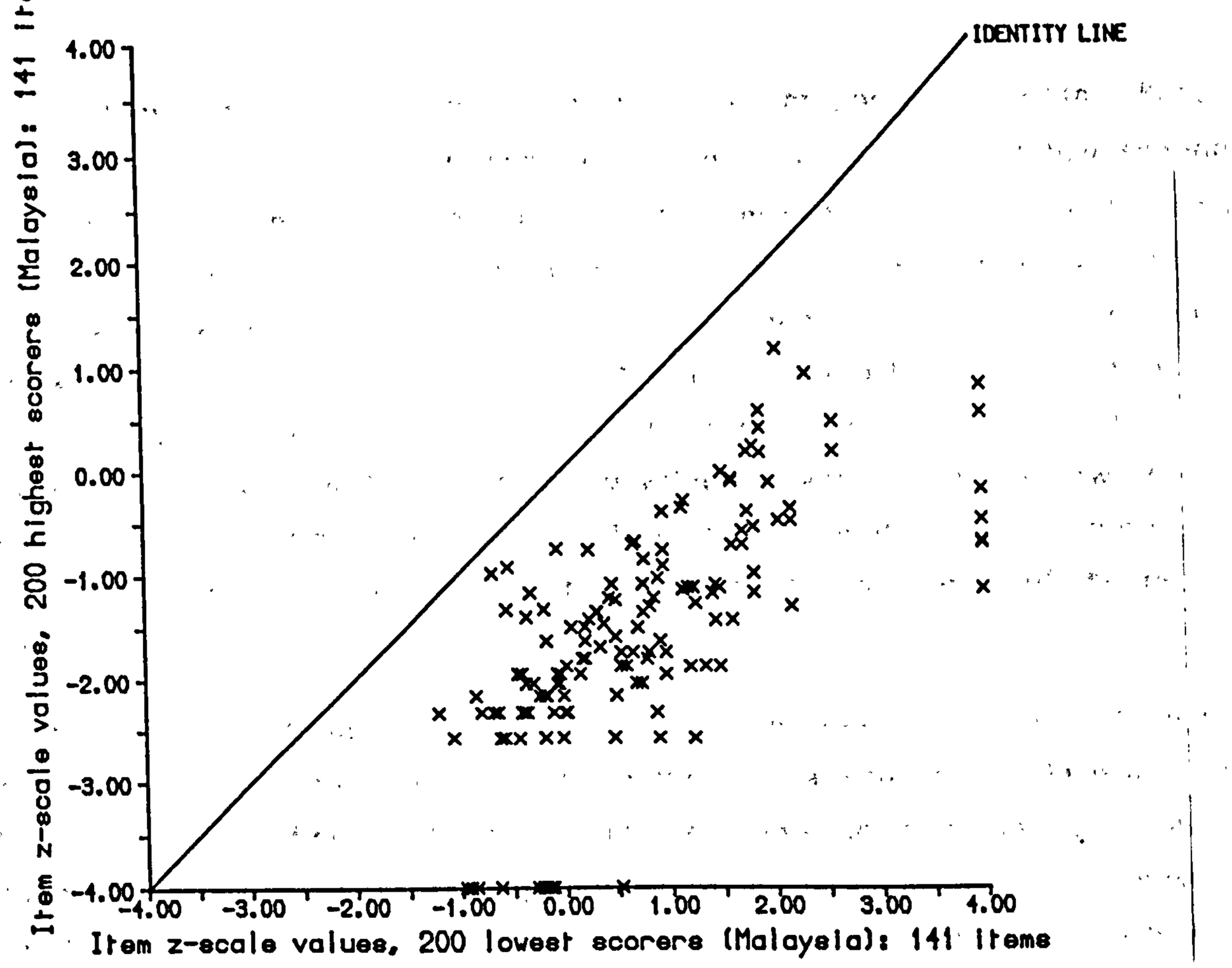


Figure 4.18 Item Z-Scale Values, High vs Low Scorers (Malaysian Data)

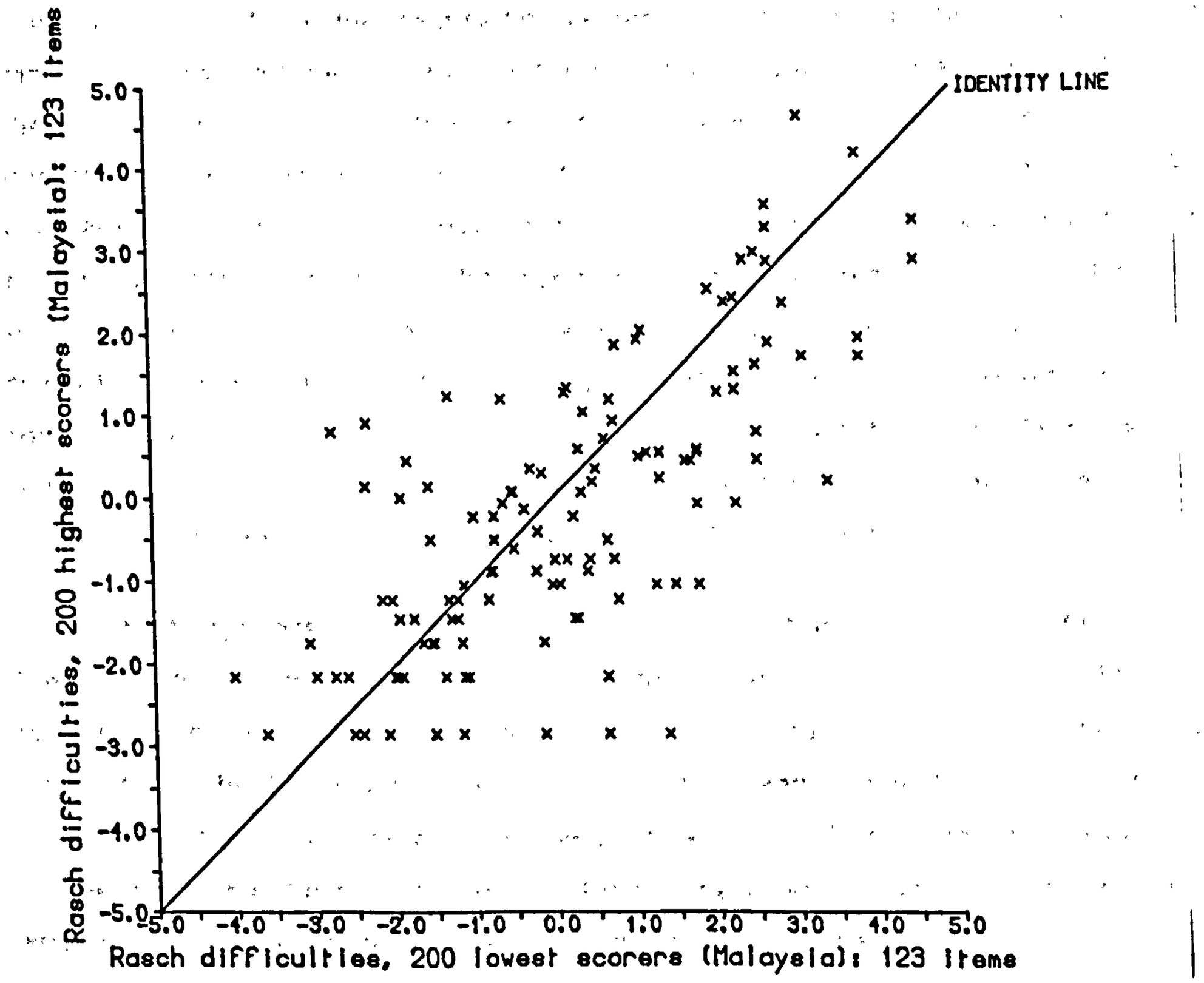


Figure 4.19 Rasch Difficulty Estimates, High vs Low Scorers (Malaysian Data)

It will be noted from Figure 4.19 that it was not possible to obtain Rasch difficulty estimates for all 141 items using only these high- and low-scoring subgroups. It was necessary to omit from their respective analyses the 11 items which were answered correctly by all members of the upper group and the 7 items which were answered incorrectly by all members of the lower group. Figure 4.19 therefore shows the pairs of Rasch difficulty estimates plotted for the 123 items which could be calibrated for both testee groups. (Since the 18 items excluded from Figure 4.19 all had facility values of either 0 for the low-scoring group or 1 for the high-scoring group, they can be identified in the plots of the traditional indices as those falling at the lower or upper limits of the ranges shown.)

It can be seen from Figure 4.19 that the Rasch difficulty estimates from the high- and low-scoring groups are, as in the comparison of the Malaysian and Tanzanian groups, free of the effects of the difference in group ability levels. Although not clustered as closely around the identity line as in the previous comparison, a fairly large number of the estimates nevertheless show a reasonable degree of consistency between groups. Again, the points are fairly well spread out along the identity line rather than being forced by a restriction of range into any particular shape. Although it may appear from Figure 4.19 that there is a lower limit at about -3.0 on the x-axis, this is not the case: this impression has resulted from the fact that easy items could not be well calibrated using the higher level group, since no information was available to distinguish them from each other for this group. The same effect can be seen elsewhere in this figure, both in the other horizontal lines formed by points towards the lower end of the x-axis and in the vertical lines formed by points towards the upper end of the y-axis; these last result from the fact that the response data of the low-scoring group contained too little information about the more difficult items for their difficulty levels to be differentiated.

Thus it appears from this comparison of results from the high- and low-scoring subgroups that the item z-scores offer a less satisfactory alternative to the Rasch difficulty estimates than the previous comparison might have suggested. The Rasch difficulty estimates have been shown to be less seriously affected by the wide difference in the proficiency levels of these subgroups, and to be preferable both in terms of stability between groups, and in not being subject to boundary effects of the kind evident in the other indices.

4.5 Rasch Analysis of Cloze-Type Data: Further Investigations

In this section, the results of a number of further investigations of the cloze-type data are presented. These are for the most part based on checks of the type proposed in the IRT literature as ways of determining (a) the extent to which particular data sets appear to meet the Rasch model assumptions, and (b) the extent to which the expected advantages of Rasch analysis are achieved in particular applications.

4.5.1 Observed vs Expected ICCs

The comparison of observed and expected item characteristic curves was mentioned in Section 4.3.2.5 in connection with the evaluation of item fit, and rough estimates of the ICCs (in the form of proportions correct across six ability subgroups) for a number of individual items were discussed.

In this section, the proportions of correct answers given to each item in the test by the six Malaysian subgroups (see Appendix E.3) are presented in graphical form, with the dual purpose of (a) further illustrating the general degree of conformity between data and model, as well as the oddity in the pattern of responses on some items, and (b) providing an immediate visual impression of the degree to which the Rasch assumption of equal discrimination is met by these data. In order to provide some examples of the corresponding expected ICCs, use has been made of the information in the 'Departure from Expected ICC' table, also in Appendix E.3, to calculate the proportion of correct answers predicted by the model for each subgroup.

For the graphs set out in Figures 4.20 to 4.23, the items have been grouped by passage, and the proportions correct plotted against the mean ability estimate of each subgroup. Although joined by spline interpolation, the individual points have been retained, to show the levels of the six subgroups in this application. The observed ICCs are shown for every item in all 12 passages (A-L). Since the expected ICCs will all be similar in shape, and will differ only in their position in relation to the x-axis (and hence in the portion that would be visible here), these are shown only for the items in the first and last two passages (A, B, K and L). Where there appear to be fewer curves than there are items in a given passage, this is simply the result of there being sets of points which were so similar as to be indistinguishable when plotted.

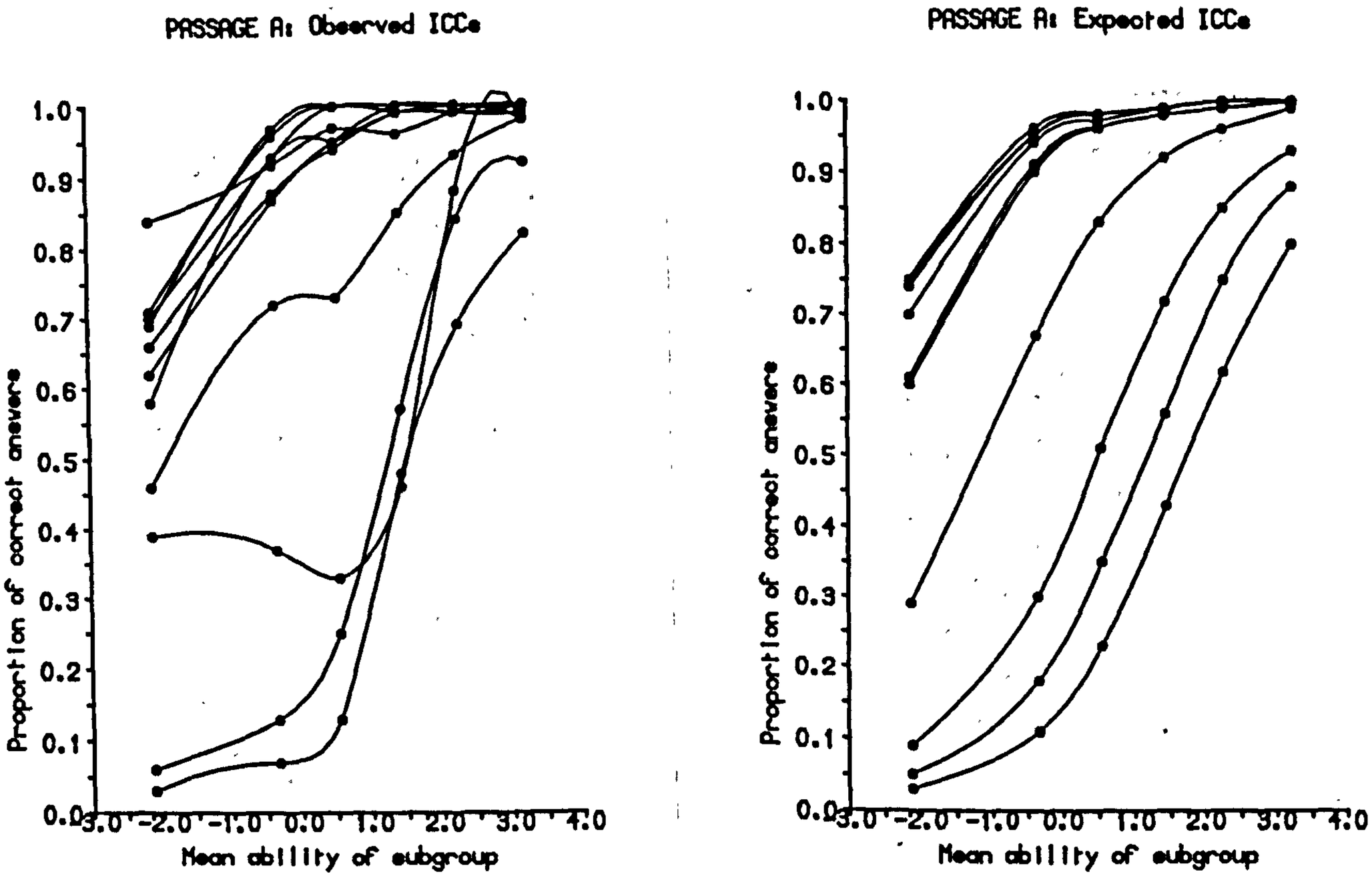


Figure 4.20 Observed vs Expected ICCs, Passages A & B

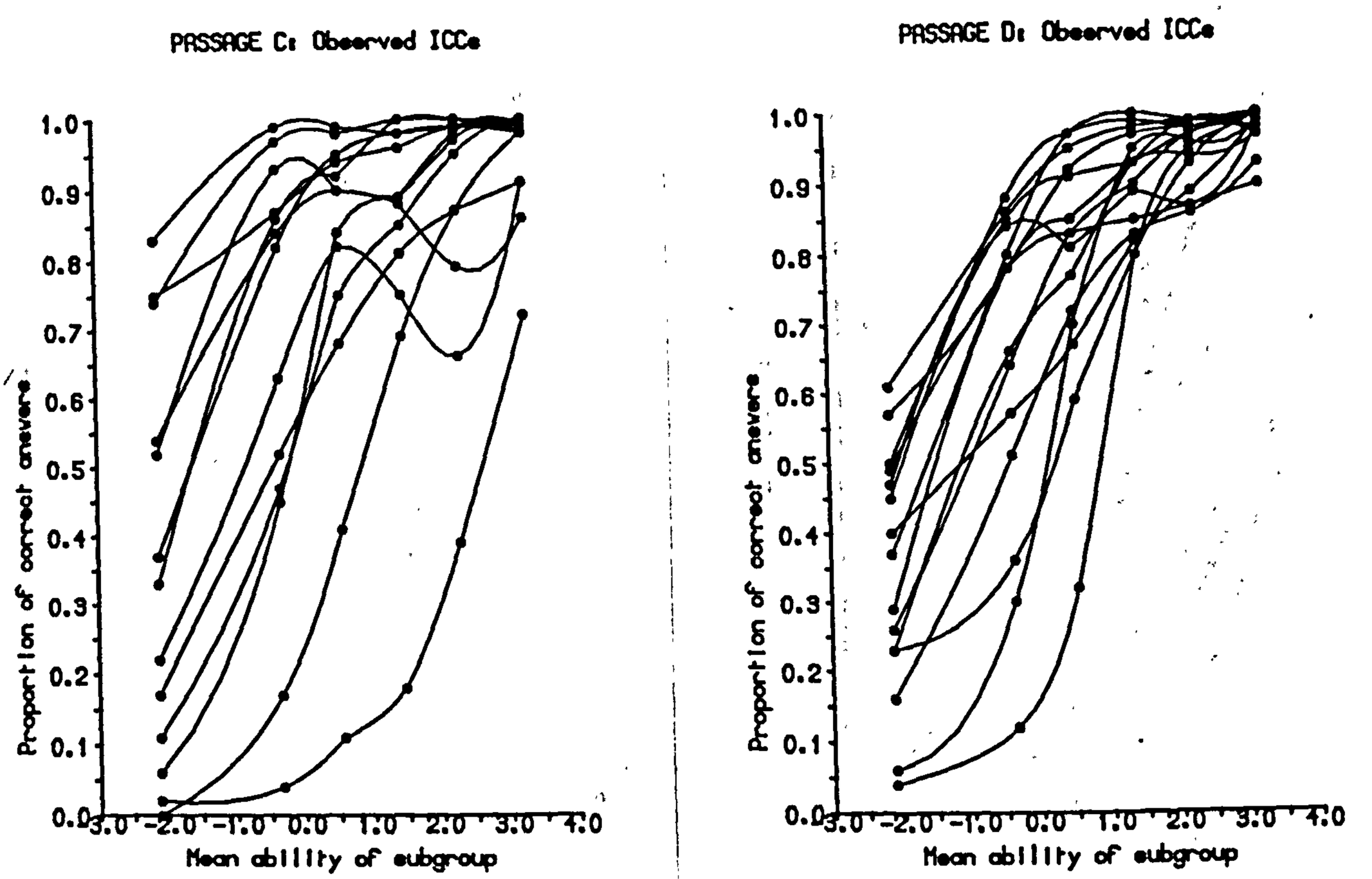
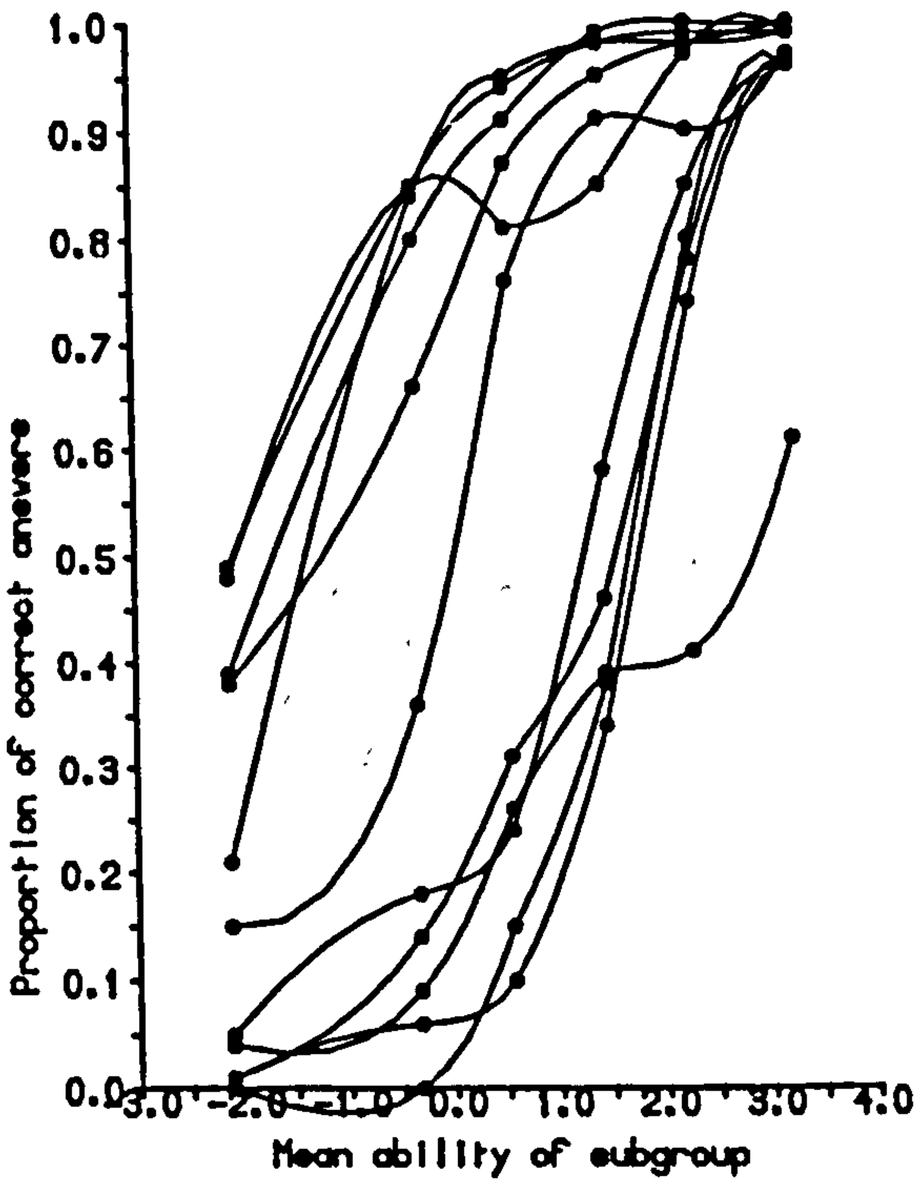
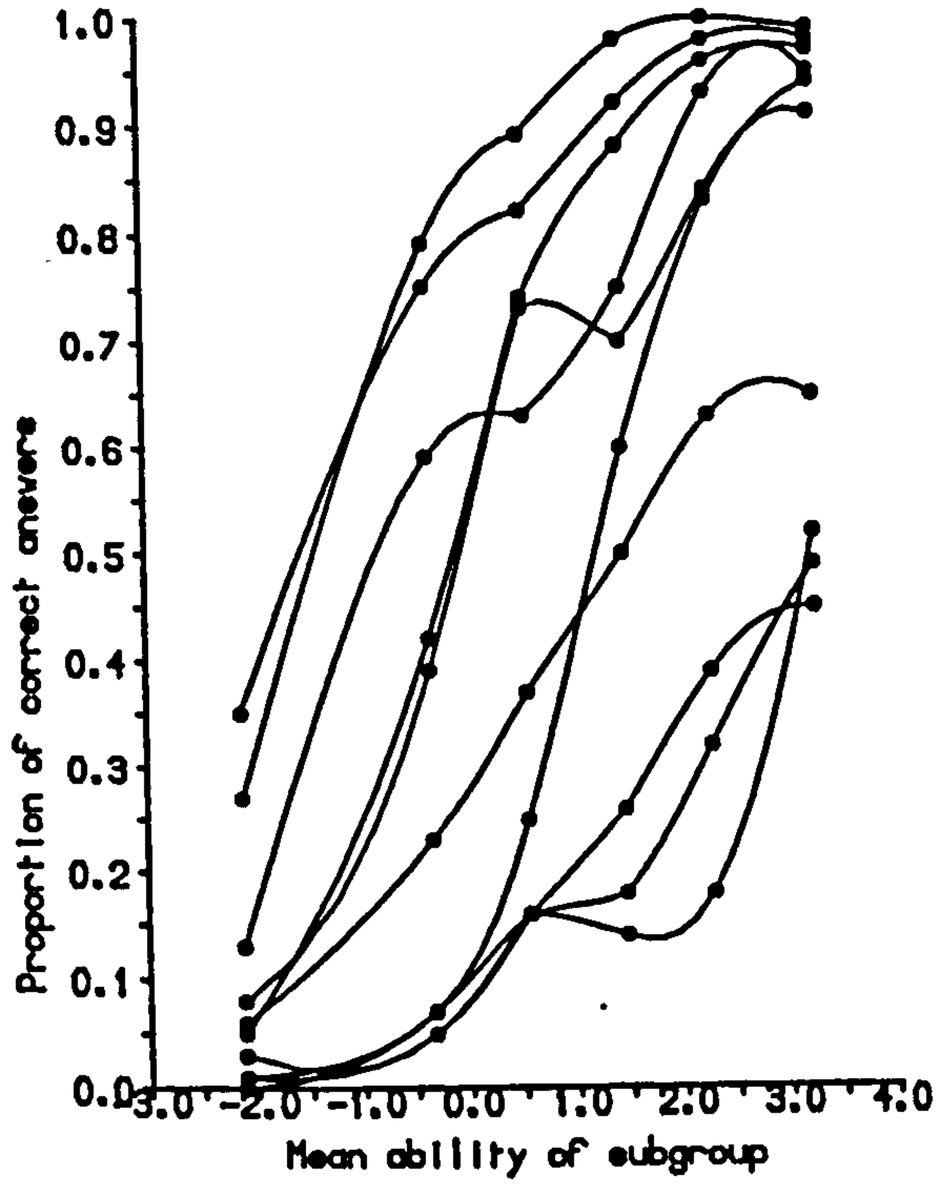
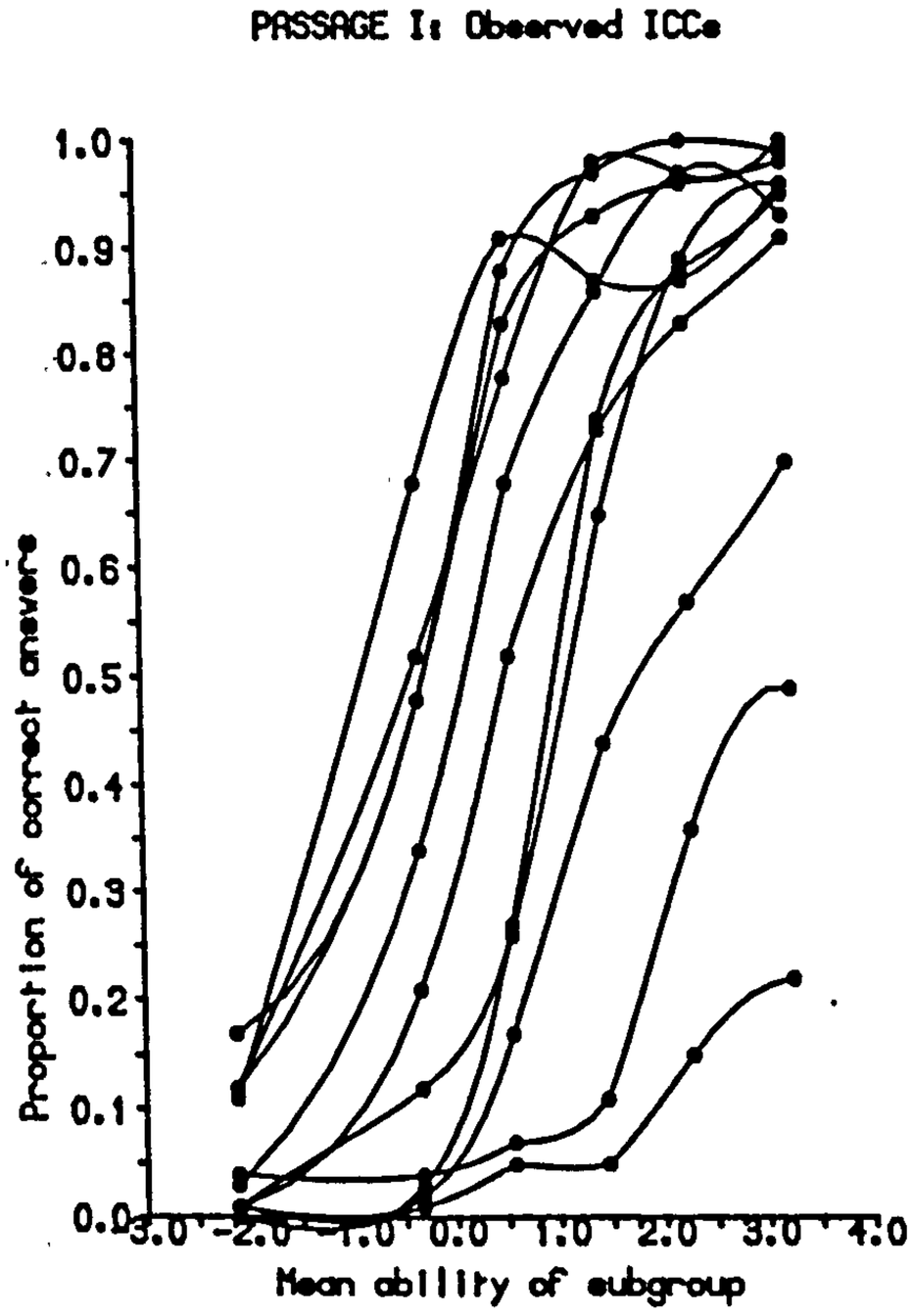


Figure 4.21 Observed ICCs, Passages C,D,E & F







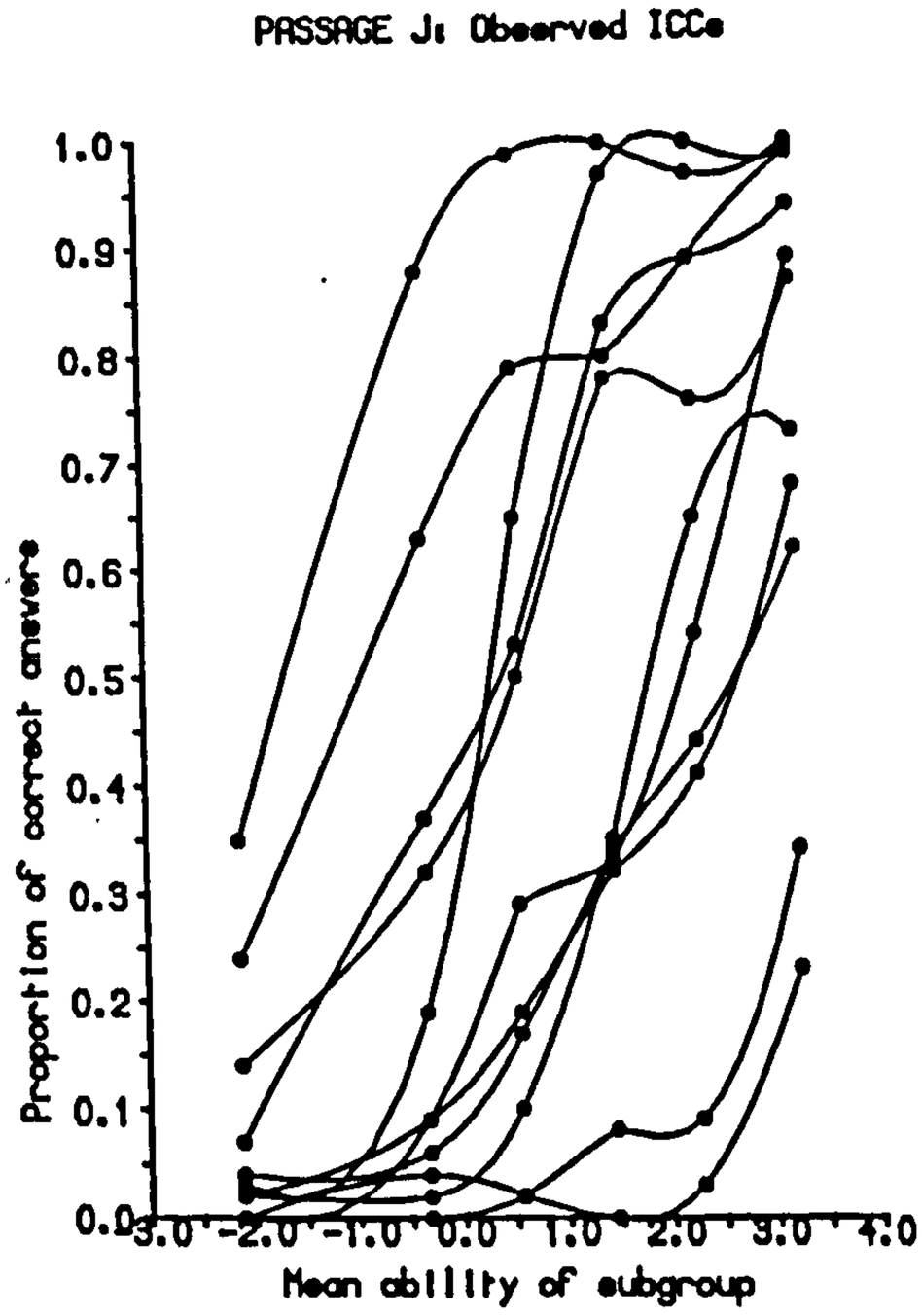
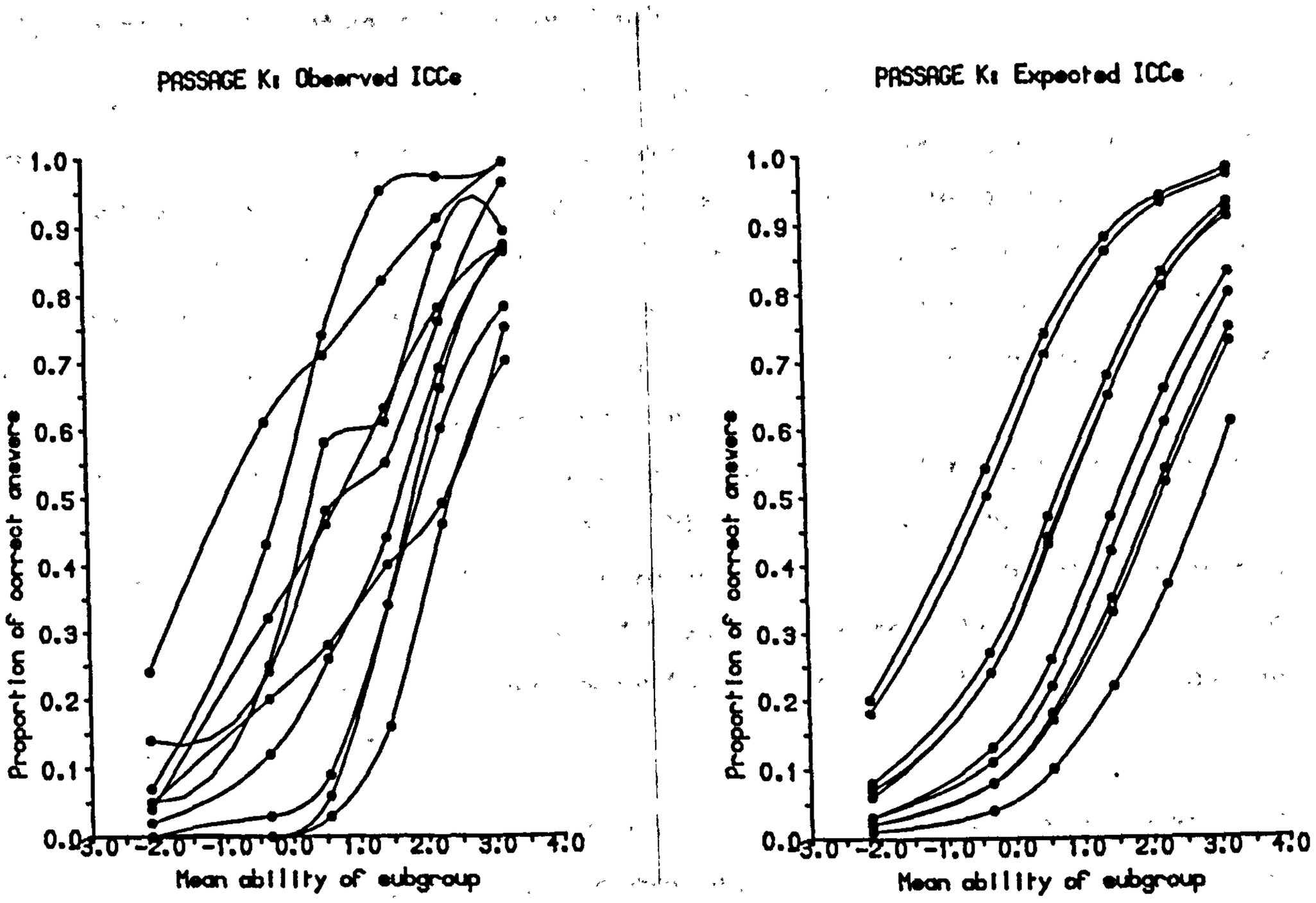


Figure 4.22 Observed ICCs, Passages, G,H,I & J



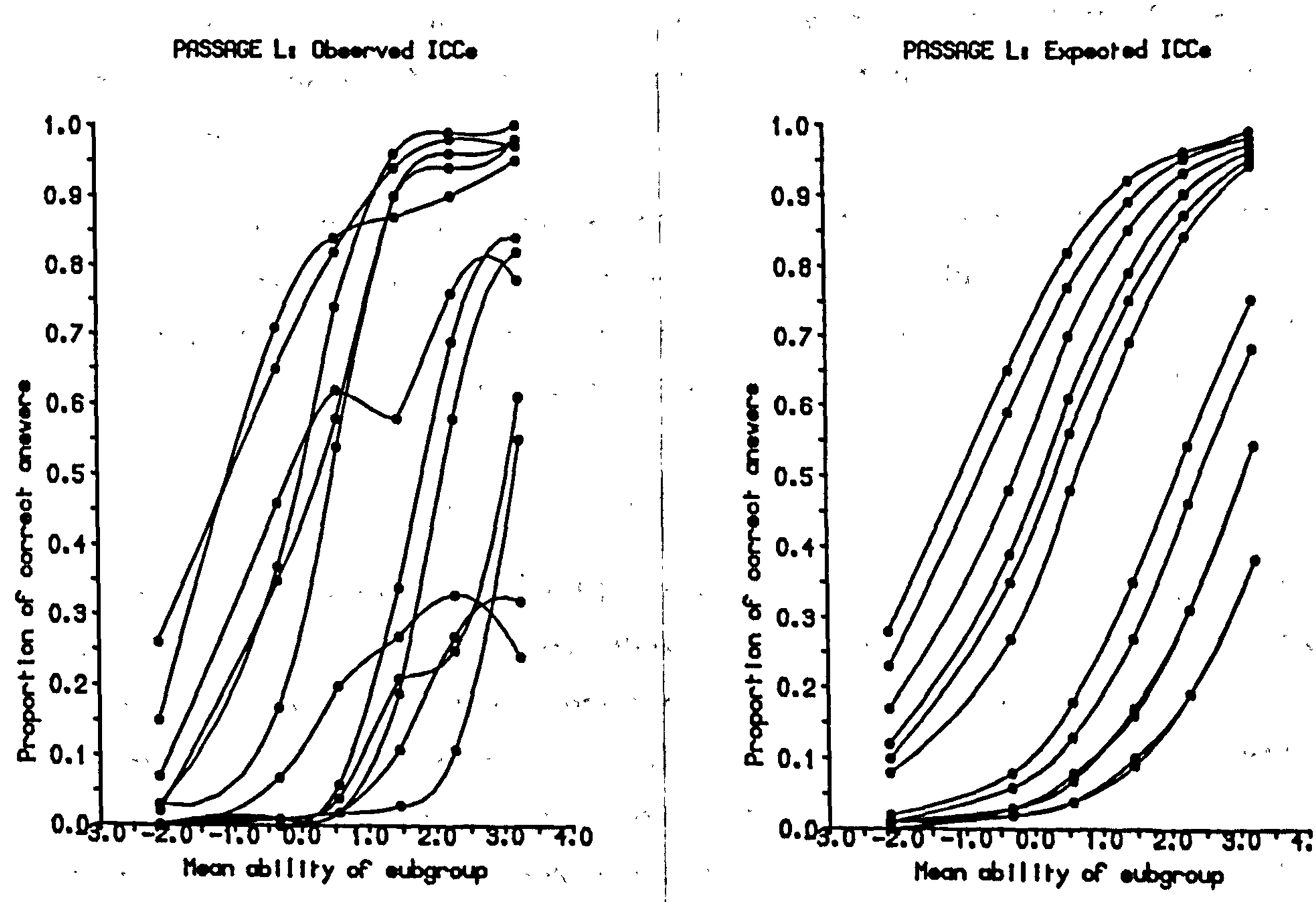


Figure 4.23 Observed vs Expected ICCs, Passages K & L

4.5.1.1 Conformity between Data and Model

Presentation of the results of the Rasch analysis in this form allows immediate identification of items for which the success rate increased as expected with overall test score, and those for which the pattern of correct answers across subgroups deviated from this; one can also see for which subgroups the greatest departures from expectation occurred. In passage A, for example, (see the first two graphs in Figure 4.20), there is for most items quite close correspondence between observed and expected ICCs; the most noticeable lack of conformity, however, appears in the observed ICC for the third item from the bottom, which begins considerably higher than expected, and then dips slightly across the 2nd and 3rd ability groups, before rising sharply across the three highest-level groups. This item (A7) was discussed in Section 4.3.2.5, since it was among the least well-fitting items for this group, both in terms of overall fit and between-group fit.

The item shown both by the traditional discrimination indices and the analysis of item fit to have given rise to the greatest inconsistency, item B13, can be easily identified in the observed ICCs for passage B shown in Figure 4.20. Although there are altogether two items in this passage for which success rate decreased at some point as mean subgroup ability increased, and one for which it remained the same, the almost v-shaped ICC for item B13 is nevertheless clearly distinguishable.

In the four graphs shown in Figure 4.21, the most noteworthy departures from a steady increase in success rate across subgroups are found among the ICCs for passages C and F, where the proportions correct can in three cases be seen to have decreased fairly sharply across the 3rd, 4th and 5th subgroups. The items for which this occurred (C34, C36 and F67) were investigated in Section 4.3.2.5, since this inconsistency was again reflected in their fit statistics.

In considering Figures 4.20 to 4.23, it should be noted that the type of curve fitted for the purpose of joining these points occasionally gives a slightly misleading impression. In the graphs for passages G and I, for example, it can be seen that the line joining the first two points where the proportion of correct answers is at its lowest descends below the x-axis in each case. This is an artefact of the plotting method used, the minimum possible value for the proportion correct being, of course, zero. A similar, though opposite, effect can sometimes be seen where proportions at other parts of the scale differ only minimally: for passage F, for example, as proportions correct approach their upper

extreme for the 5th and 6th subgroups, the ICCs appear, artificially, to rise and fall between these two points.

Bearing in mind, then, that genuine changes in direction of the ICC occur only where the plotted points change in direction, Figures 4.20 to 4.23 indicate that a large majority of the ICCs show a reasonable approximation to the expected shape, but that most passages contain at least one, and often two, items whose ICCs decrease noticeably at some point as mean estimated ability for the subgroup increases. As has been indicated, attention will normally be drawn to such items by the item fit statistics, in particular the between-group fit t-value.

In the light of the suggestions made in Section 4.3.2.5 as to the possible reasons for some of the observed inconsistencies, it would be of interest to re-mark the test papers according to an amended marking scheme, to re-analyse the data, and to plot the resulting item characteristic curves. One would then be able to see whether changes of the type proposed on common sense grounds as being likely to improve the measures yielded by this test would in fact result in new ICCs approximating more closely to the theoretical ICCs predicted by the Rasch model.

4.5.1.2 Assumption of Equal Discrimination

It is clear from Figures 4.20 to 4.23 that there is variation in the discriminating power of these items for this group, i.e. that the observed ICCs differ in their steepness. The expected ICCs, by contrast, would, if shown across a wider ability range, be approximately uniform in shape.

The Rasch-based discrimination index included among the fit statistics in Appendices E.4 and F.4 reflects the difference in steepness between observed and expected ICCs. As Wright et al. (1980:84-85) explain, values close to 1 indicate that the observed and expected ICCs correspond closely. A value of considerably less than 1 indicates that the ICC is flatter than expected, and hence that the item discriminates less well than the other items in general, while a value of considerably greater than 1 indicates that the ICC is steeper than expected, and hence that the item differentiates more sharply than average among the different ability levels.

The Rasch-based discrimination indices for the main (Malaysian) data set considered here can be found in Appendix E.4. The values range from -0.02 (for item B13) to 1.5 (for item J116). A summary of the distribution of values for the

complete item set is given below; this provides an indication both of the amount of variation in discriminating power observed here and of the nature of the departures from the expected steepness of slope.

Rasch-based discrim. index	No. of items
.39 or less	3
.4 to .69	12
.7 to .99	46
1.0 to 1.29	62
1.3 or over	18

It can be seen from this summary that for more than 100 of the 141 items, the observed and expected ICCs corresponded quite closely in steepness. Of the remainder, just over half were steeper than expected, and the rest flatter. The information provided by this index confirms the general impression given by the appearance of the curves plotted in Figures 4.20 to 4.23, i.e. that while many of them rise steadily across the ability range, a small number in each passage deviate from this.

Of course, one would never expect all items to show the same discriminating power in any real application; the question is rather that of whether reasonable conformity to some sensible model of performance is evident. Although, as was mentioned in Chapter 2, it is sometimes suggested that failure of the data to meet this (or any other) assumption of the Rasch model should be taken as a warning that Rasch analysis is inappropriate for use with that data set, this view overlooks the important point that it is precisely in analysing the data in accordance with such a model that one gains useful information about the test. Again, it would be of interest to compare the Rasch-based discrimination indices obtained in this analysis with those from a re-analysis after revisions to the test procedure in order to observe any changes in their distribution as a result of 'improving' the test.

4.5.2 Dimensionality of the Data

The assumption of unidimensionality, although made implicitly and apparently without question by those who use traditional methods of test analysis, has given rise to a great deal of discussion in the IRT literature, and it has again been suggested that data sets should be investigated from this point of view before applying methods of analysis deriving from unidimensional item response models. Although this is not the approach taken here (since the results of such analyses can themselves be seen to have yielded useful information about the

test) it is nevertheless of interest to examine the main data set used here for evidence of violations of the unidimensionality assumption, by means of some of the methods suggested in the literature.

It should be remembered that the concern in these investigations is with evidence that this cloze-type test tapped two or more poorly correlated dimensions, and that it is not the purpose of this study to analyse the construct underlying performance on this test. Although one might suggest a number of different abilities that might come into play in answering these items, it is only if these appeared to be generally poorly correlated within the target population that the unidimensionality of the data would be called into question.

4.5.2.1 Guessing and Time Effects

In a constructed response test such as this, it seems extremely unlikely that testees will be able to supply correct answers completely by chance. Indeed, the selection of answers without regard for context is likely to stand out clearly, as it does in some of the response patterns observed in the Malaysian data set. Two strategies noted where testees appeared not to have used the context provided to complete the blanks were (i) the use of words taken seemingly at random from other parts of the test, and (ii) the choice of e.g. the definite article for a large number of the blanks, presumably in the hope that this would be correct at least some of the time. Not surprisingly, neither approach proved fruitful; even the second strategy, which perhaps showed greater test-wiseness, was unsuccessful since, as can be seen from the marking sheet in Appendix B.2, there is no single answer which occurs particularly frequently. Thus as far as this test is concerned, guessing can be discounted as a serious threat to unidimensionality.

The possible effects of the time limit, on the other hand, merit closer investigation (insofar as this is possible using the response data alone), to see whether the test appears to have confounded separate dimensions of proficiency and speed. The main evidence available for this purpose is in the number of persons omitting each item. An accumulation of omissions towards the end of the test would tend to suggest that candidates had run out of time; it should be borne in mind, however, that where items are ordered by difficulty, some omissions of later items are likely to occur for this reason alone, particularly if the test is not one which permits guesses to be made rapidly and without effort.

Examination of the numbers of Malaysian testees omitting each item in this

test reveals that there were omissions throughout, and not only towards the end; indeed, no item in the test was attempted by all 611 testees, and even in the first three passages there were ten items which were left blank by at least 20 people, one of these (C27) having been omitted by as many as 68 people. Lack of time, then, was clearly not the only reason for omissions in this case. Although there were generally more omissions in the later part of the test than in the earlier part, the numbers do not show a steady accumulation from middle to end. There were, for example, noticeably fewer omissions of items in passage J than in passage I, suggesting that it was the difficulty of the items, and not merely their position in the test, which caused people to omit them. Furthermore, even the last two items in the test were omitted by considerably fewer candidates (74 and 88 respectively) than some of the earlier items.

Thus although it would appear from the generally larger numbers of omissions in the latter part of the test that the time limit had at least some effect, it also seems likely that the greater difficulty of the later passages (if not necessarily of the individual items within them) will have deterred some of the lower-level candidates from attempting answers.

In view of the time allowed for completion of the test, the simplicity of some of the earlier passages, and the sharp discrimination noted in connection with some of the items occurring near the end of the test, there seems little reason to suppose that high-level testees would have been prevented by lack of time from attempting all items. Although in order to be certain of this one would need to examine the numbers and positions of items omitted by high-scoring persons, it appears unlikely that speededness has interfered in any serious way with the measures yielded by this test.

4.5.2.2 Division of Data by Item Subsets: Comparison of Difficulty Estimates

One of the suggested methods for investigating the dimensionality of test data is that of Bejar (1980), in which item parameter estimates are obtained separately for subsets of items which it is thought might constitute separate content dimensions, and compared with estimates obtained from the item set treated as a whole. This method was outlined in Chapter 2, where it was also noted that the rationale for this procedure has recently been questioned (by Spurling, 1987). Given the disagreement concerning the validity of this method, its use is demonstrated here, and the results compared with those obtained using an alternative procedure considered by Spurling to be acceptable (see Section 4.5.2.3).

One of the difficulties in applying these methods to data from a cloze-type test is that there is no obvious way in which to divide the items into subsets which might be thought to tap separate, uncorrelated abilities (in the case of the ELTS data analysed in Chapter 5, on the other hand, an obvious division by subtest is possible). Although, as was mentioned in Chapter 3, a number of researchers have sought to identify the different abilities involved in answering cloze items, the results of these studies are not necessarily applicable for this purpose. For example, the tendency noted by Bachman (1985) for the difficulty of cloze items to increase with the amount of context required for their completion, though interesting in itself, does not carry any implication that the ability to answer e.g. items requiring only the context provided by the same clause is uncorrelated with the ability to answer items depending on wider-ranging context.

In the study by Lee (1985), however, it is suggested that two abilities, corresponding to an 'openness' vs 'closedness' contrast, might underlie performance on cloze items: 'open' items would be those having a range of possible answers, while 'closed' items would be those for which the possible fillers were very restricted in number. Since Lee used the exact word scoring method, 'open' items were characterised as being less predictable than 'closed' items. Although this would not necessarily be the case for the cloze-type test under discussion here, for which a form of acceptable word scoring was used, it was decided that the general notion of an 'open' vs 'closed' contrast could nevertheless be applied here.

For the purposes of this part of the study, then, the items were grouped in two different ways: (i) by the familiar content vs structure word division, and (ii) by an 'open' vs 'closed' division based on the number of different answers accepted as correct: items for which the marking sheet allowed more than one possible answer were classified as 'open', while those for which only one answer was accepted were classified as 'closed'. It is not hypothesised here that either of these oppositions represents a content dimension division; the purpose in reporting these investigations is rather (a) to consider the information obtained using Bejar's (1980) method and to compare it with that obtained using an alternative method, and (b) to provide a point of comparison for the investigations presented in Chapter 5, in which the same methods are applied to a different type of test.

The particular items falling into the two different categories in each case can

be seen from Appendices G.2 and G.3, which list the Rasch difficulty estimates obtained for the Malaysian data set from separate calibrations of the four item subtests created by the divisions described above.

Following the method suggested by Bejar, the difficulty estimates for each subset treated separately were plotted against the final difficulty estimates obtained for the same items from the original calibration of the complete item set (listed in Appendix E.2). The resultant graphs for the content and structure word item subsets are shown in Figures 4.24 and 4.25.

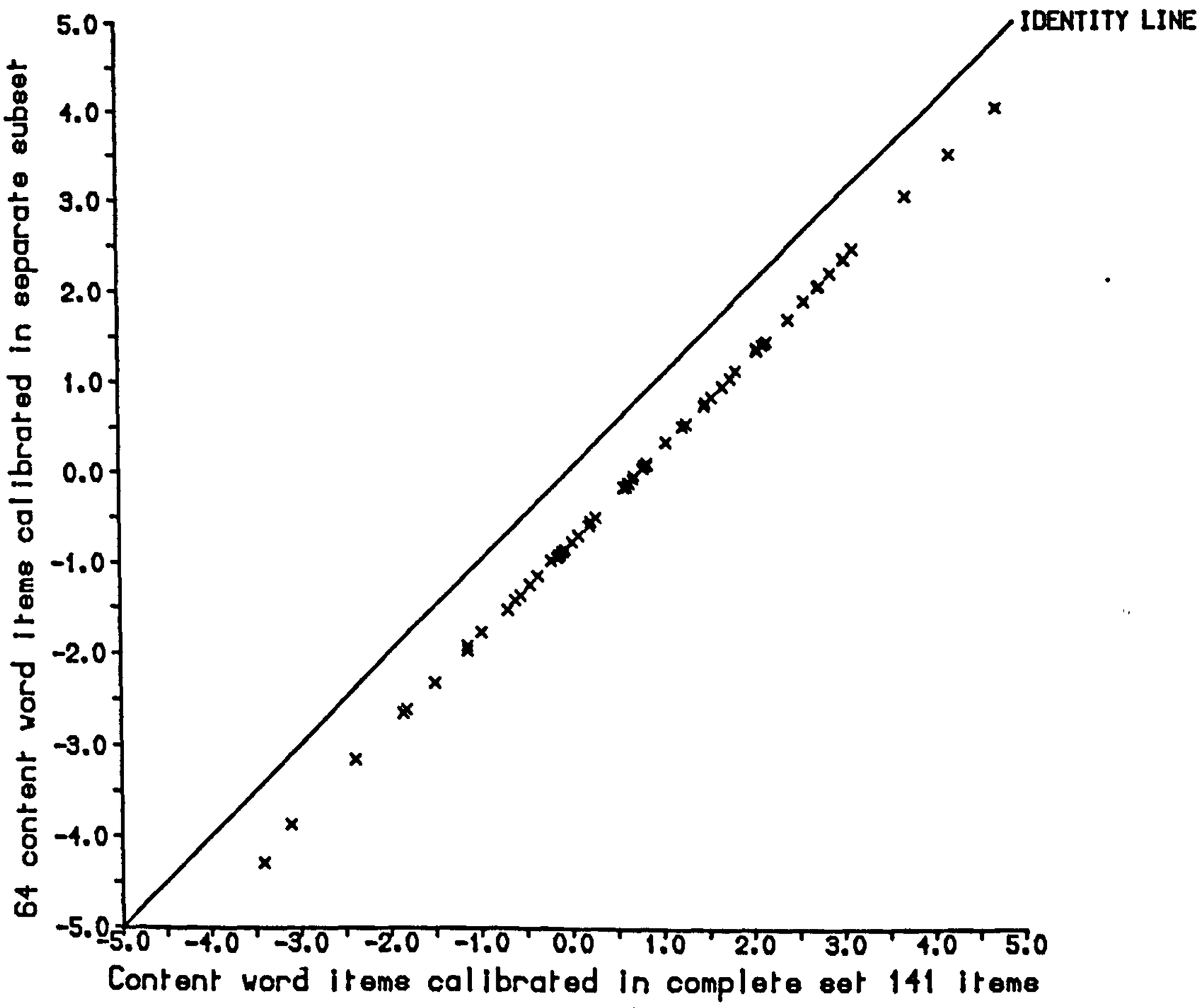


Figure 4.24 Subset- vs Test-Based Difficulties, Content Word Items

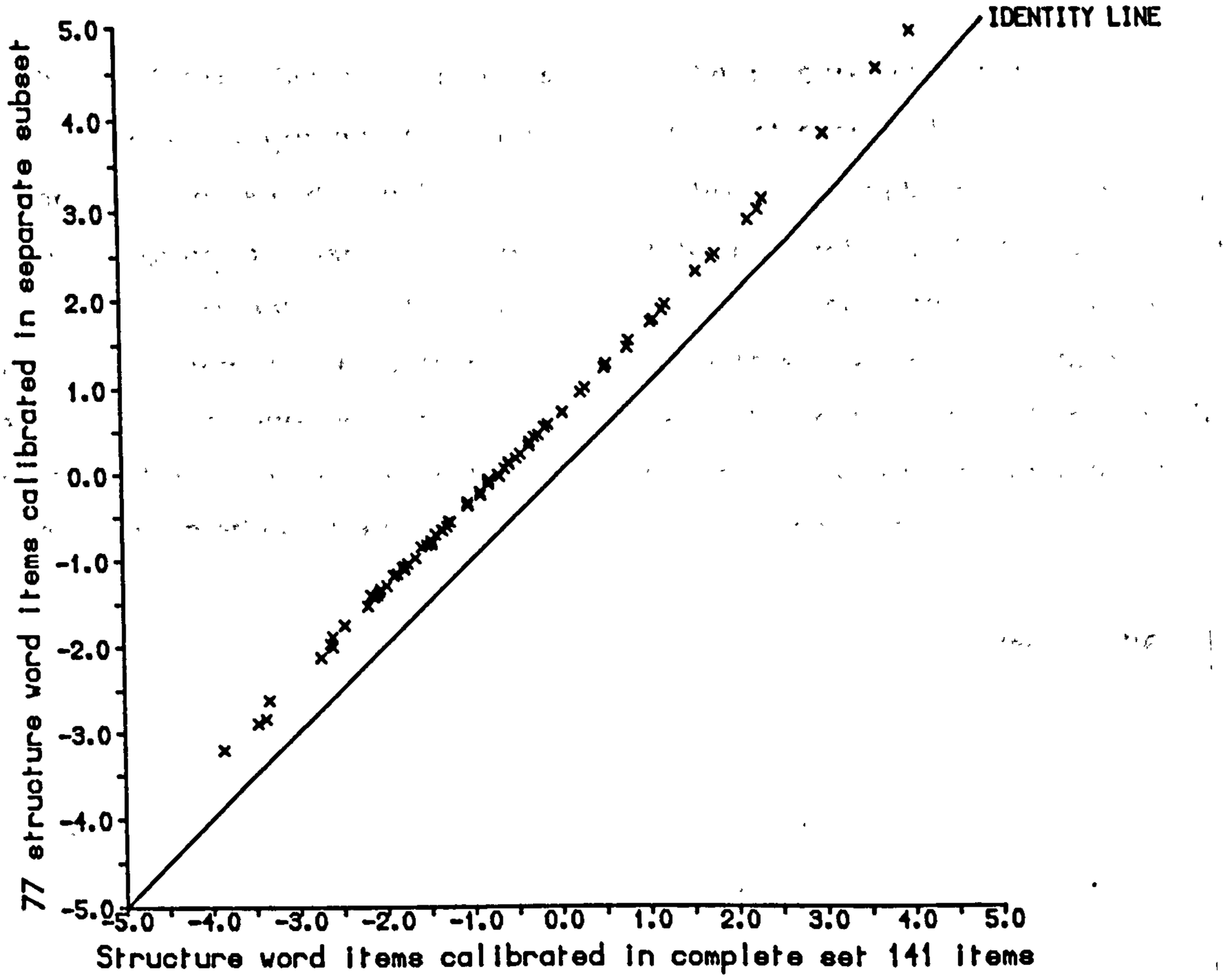


Figure 4.25 Subset- vs Test-Based Difficulties, Structure Word Items

It can be seen from Figures 4.24 and 4.25 that both for the content word and the structure word items, the plotted points form a line parallel with the identity line except at the extremes, where the standard errors will be at their largest. Since no adjustment has been made to make the mean item difficulties for the subset-based calibrations coincide with those for the total test-based calibrations, the graphs also show that when treated as part of the complete item set, the content word items are, on average, among the more difficult items in the test, while the structure word items are, on average, among the easier items. Had the means of the subset-based item calibrations been adjusted to coincide with those of the total test-based calibrations, the points would in each case lie along the identity line (apart from the slight discrepancies at the extremes).

The results are similar for the 'open' vs 'closed' division, for which the plots of the subset-based difficulty estimates against the total test-based difficulty estimates are shown in Figures 4.26 and 4.27. In this case it can be seen that the 'open' items were, on average, more difficult than the 'closed' items. Again, had this effect been removed by adjusting the mean difficulties for each subset, the points would lie along (or at least extremely close to) the identity line.

As regards the interpretation of these results in terms of the dimensionality of the data, the conclusion one would reach following Bejar's reasoning would be

that neither of the divisions of items made here represents a real content division. Spurling's argument, however, would be that even if these divisions did correspond to separate dimensions, results of this kind would not indicate this, since the information used for the subset-based and the total test-based calibrations is essentially the same, in that it consists of the same number-correct item scores obtained from the same set of testees. The method suggested by Spurling as offering a more satisfactory alternative to the comparison of sets of difficulty estimates is the comparison of sets of ability estimates. This method is applied in the section which follows.

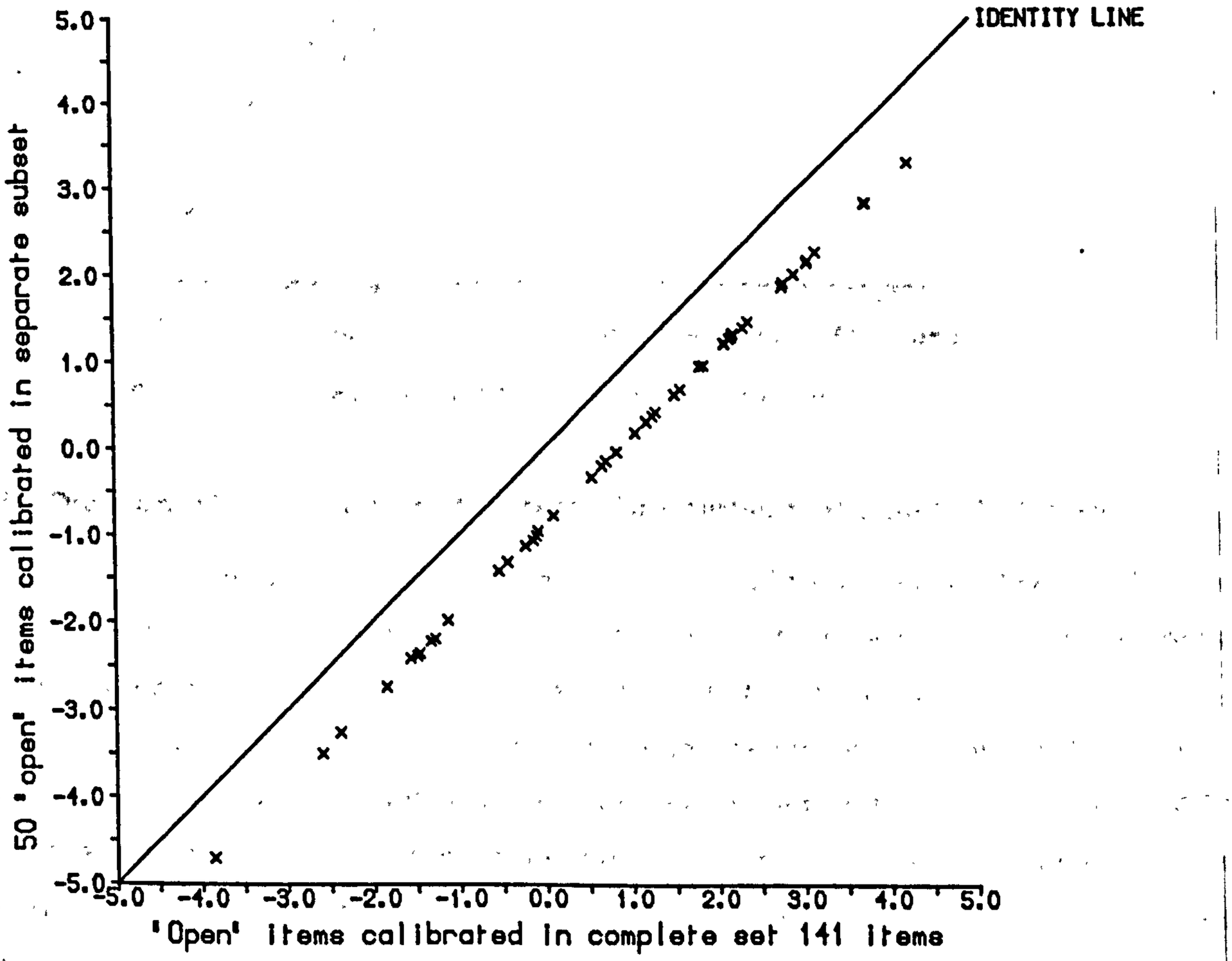


Figure 4.26 Subset- vs Test-Based Difficulties, 'Open' Items

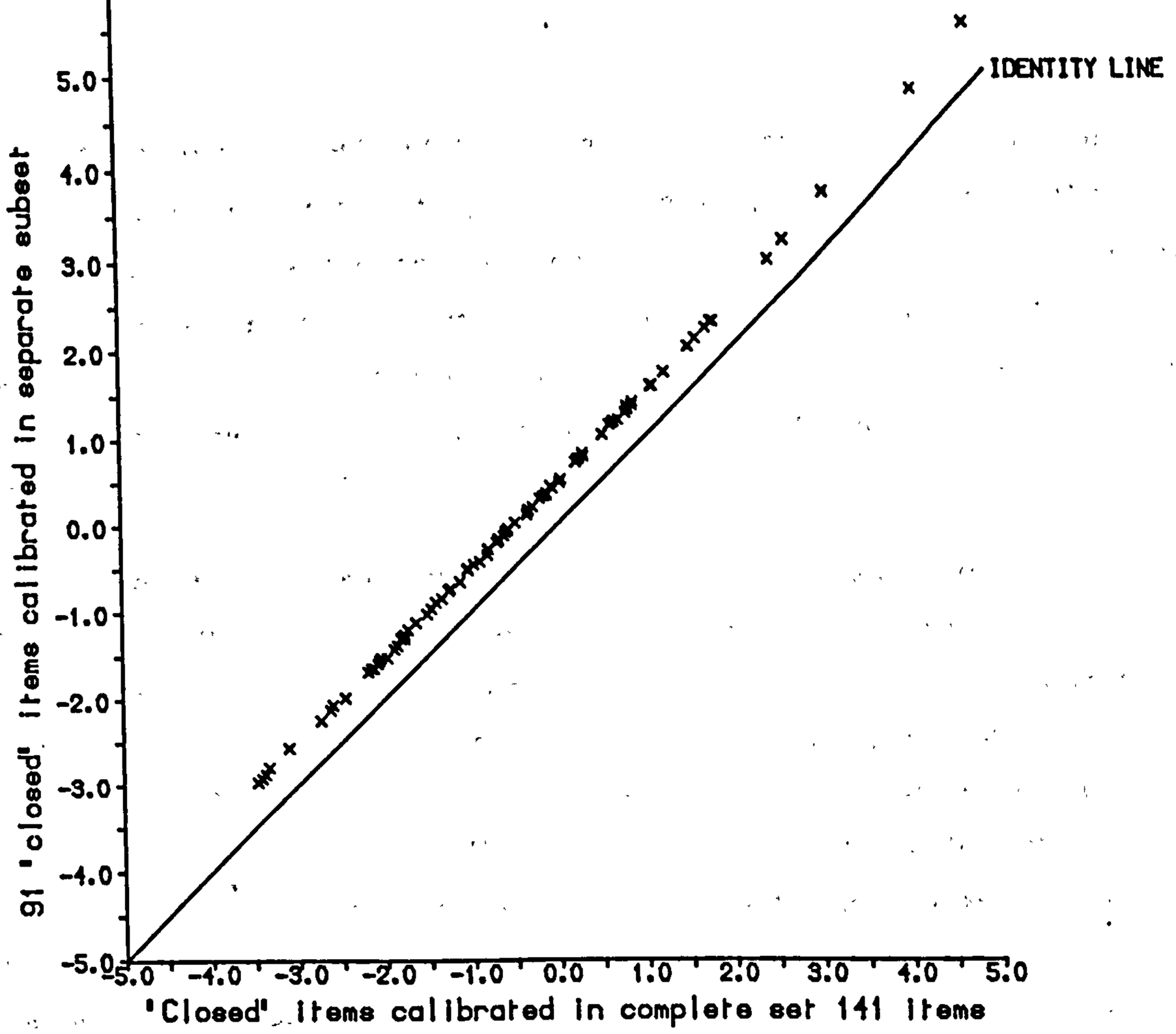


Figure 4.27 Subset- vs Test-Based Difficulties, 'Closed' Items

4.5.2.3 Division of Data by Item Subsets: Comparison of Ability Estimates

Using the same two divisions of items, ability estimates were obtained for the Malaysian testees from each of the four item subsets treated as separate tests. For all the persons remaining after the removal of those with zero or perfect scores, the pairs of ability estimates obtained using each item subset were plotted against those obtained using the complete item set. Figures 4.28 to 4.31 show the results for the 607 persons remaining in the analyses based on the content vs structure word division, and for the 606 persons in the analyses based on the 'open' vs 'closed' item division.

The rationale for this approach to the investigation of the dimensionality of test data is that if the various item subsets tap the same ability, the pairs of ability estimates for each person will show a high positive correlation: indeed, if the mean item difficulty for the subset calibration is set to that of the same items in their whole-test calibration, the pairs of ability estimates should be (approximately) the same. If, on the other hand, the different item sets measure uncorrelated abilities, the separate ability estimates will be functions of the particular item sets upon which they are based, and, when plotted, will to this extent depart from a straight line.

In the analyses carried out here, the mean item difficulties for the different item subsets were not adjusted, and thus one would not require the pairs of ability estimates to be the same in order for the assumption of unidimensionality (at least with respect to the particular divisions made) to be upheld. One would, however, require high positive correlations between the sets of ability estimates, and these, as may be seen from Figures 4.28 to 4.31, are observed in each case here.

It will be noted, however, that there is in each of these graphs an area of the scale in which the points are more widely spaced, indicating greater variability in the estimates. For the content word items and the 'open' items this occurs at the lower end of the ability scale (see Figures 4.28 and 4.30), while for the structure word items and the 'closed' items it is observed at the upper end of the scale (see Figures 4.29 and 4.31). This may be explained by reference to the difference in difficulty between the two item sets formed by each division of the data: in the case of the two harder item subsets (content word and 'open'), there appears to have been a 'floor' effect, so that low-level testees whose abilities were differentiated when measured on the whole item set did fairly uniformly badly when measured using these subsets, and thus appear on the basis of the subsets to have the same abilities. For the two easier item subsets, a 'ceiling' effect is evident at the upper extreme of the ability range. Again, testees whose abilities were differentiated by the complete item set could not be differentiated using these more restricted (and, in terms of difficulty levels, less appropriate) item sets.

The increased vertical distances observed between points at both ends of the ability scale in all four graphs result from the fact that the differences in Rasch ability estimates corresponding to differences of one raw score point are greater at the extremes of the scale than nearer the centre.

For the most part, however, the pairs of estimates in the four graphs cluster quite closely along a straight line, indicating that the separate item subsets created here do not measure separate abilities uncorrelated with that/those measured by the whole test. The conclusion, then, is in this case the same as that reached using Bejar's method. This is not to say, however, that the two methods are equally appropriate: although one would need to perform checks of this type using data sets known to confound poorly correlated dimensions in order to be able to demonstrate the alleged flaw in Bejar's method, the approach based on the comparison of ability estimates is to be preferred on logical

grounds.

A more stringent check on whether different item subsets tapped different abilities would be the direct comparison of ability estimates obtained using those subsets. Unlike the examples presented in this section, the sets of items for which abilities were compared would contain no items in common. Such comparisons were carried out in this study using the item divisions already described; discussion of these is, however, deferred until Section 4.5.4, when the test-independence of ability estimates is considered.

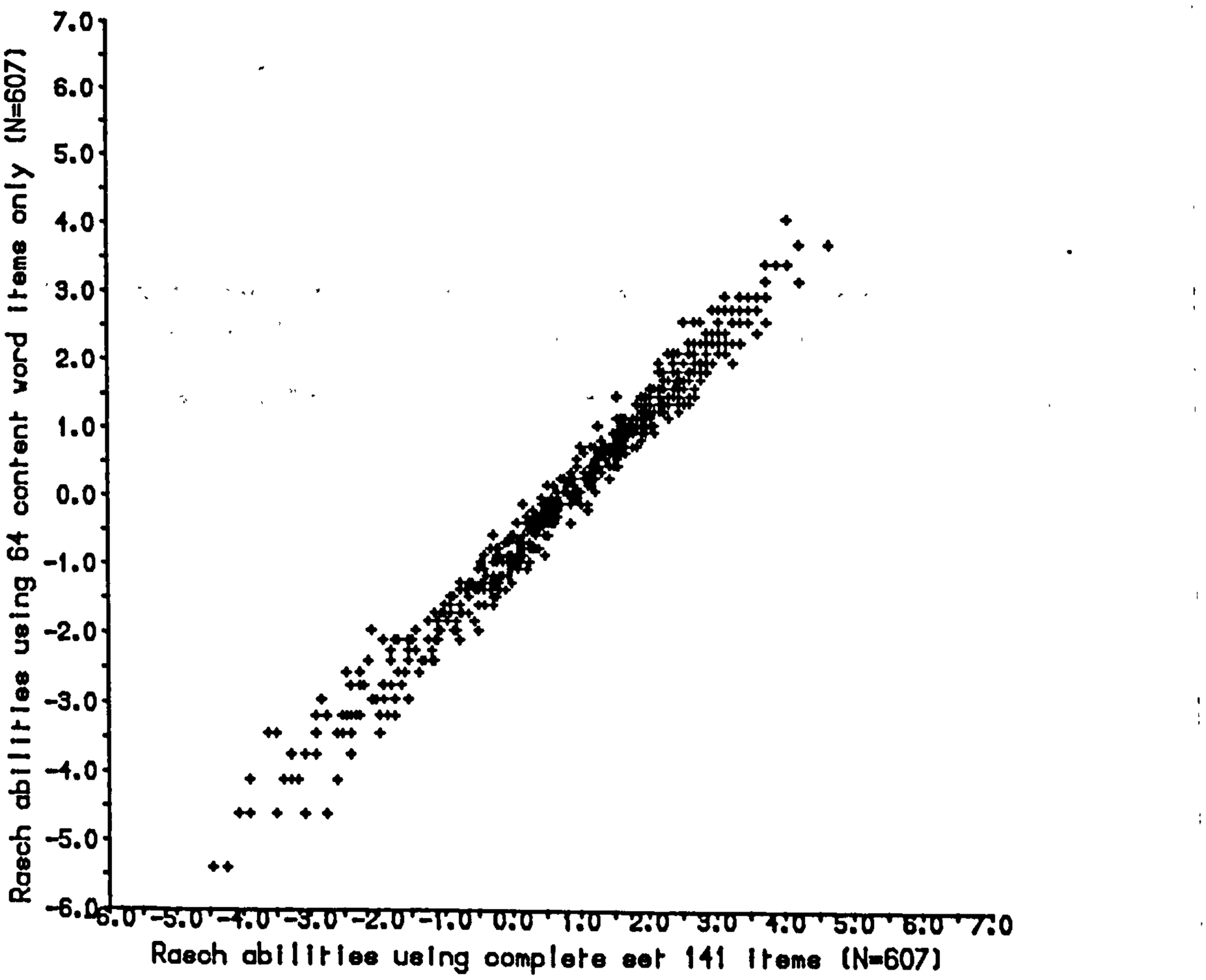


Figure 4.28 Subset- vs Test-Based Abilities Using Content Word Items

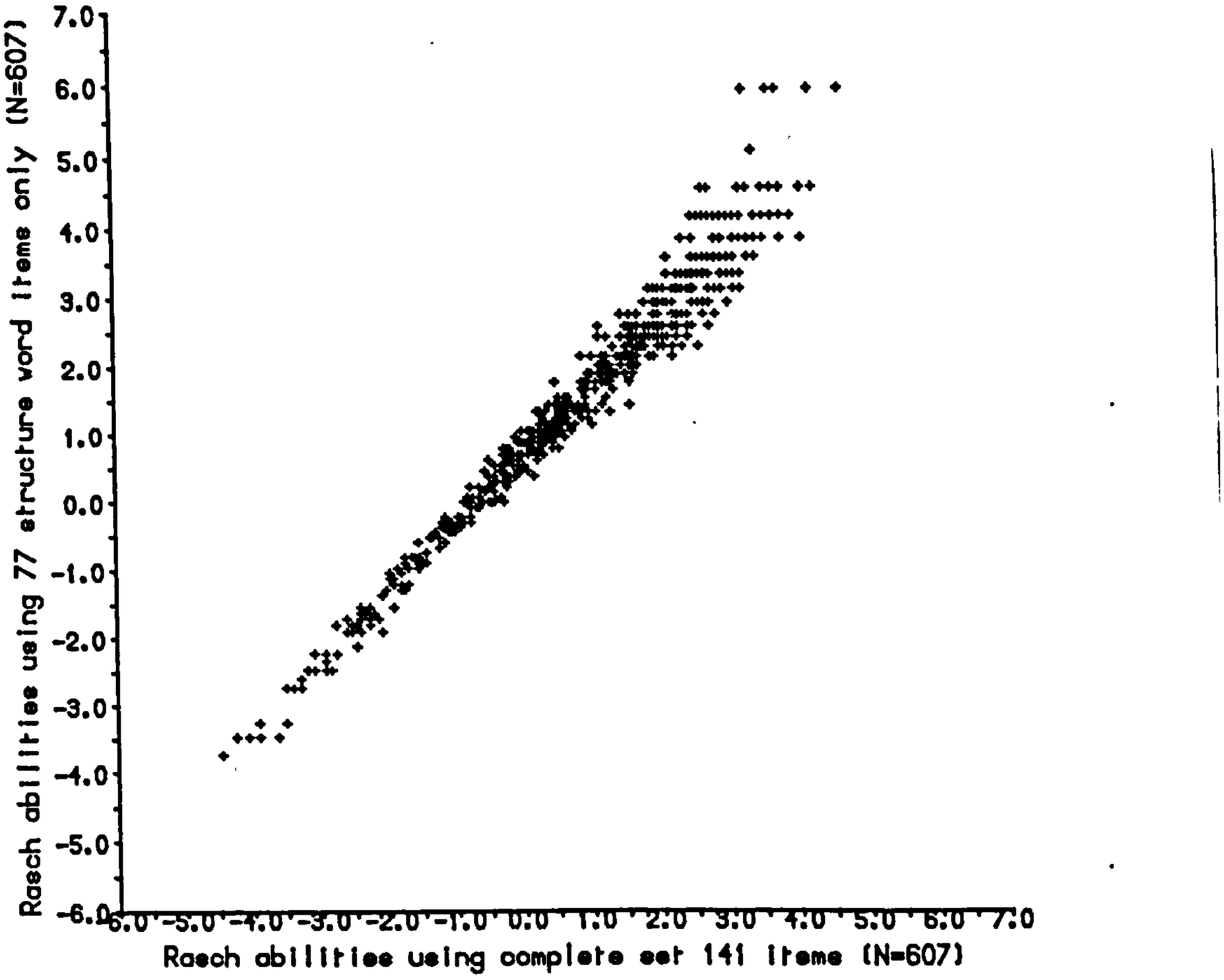


Figure 4.29 Subset- vs Test-Based Abilities Using Structure Word Items

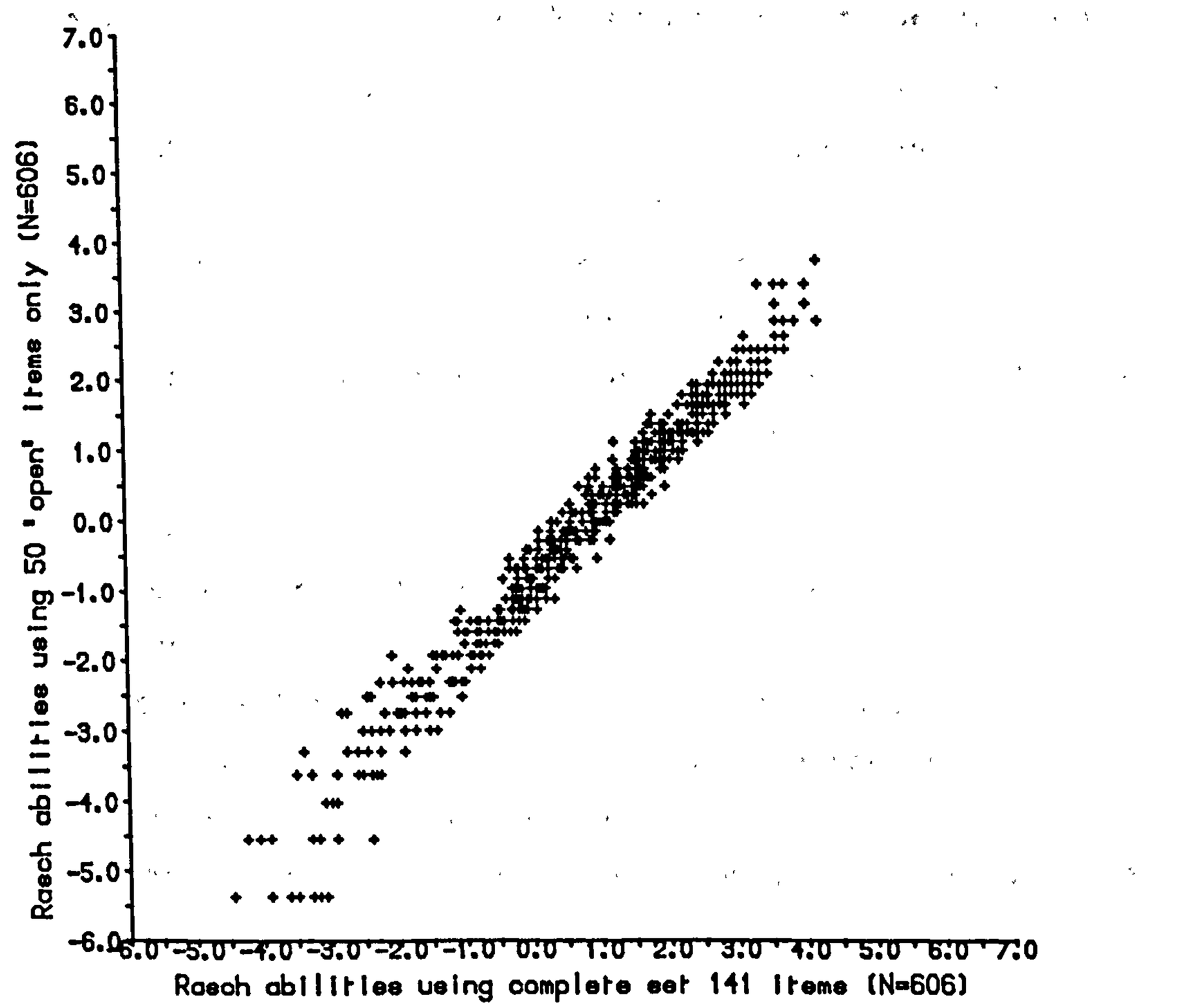


Figure 4.30 Subset- vs Test-Based Abilities Using 'Open' Items

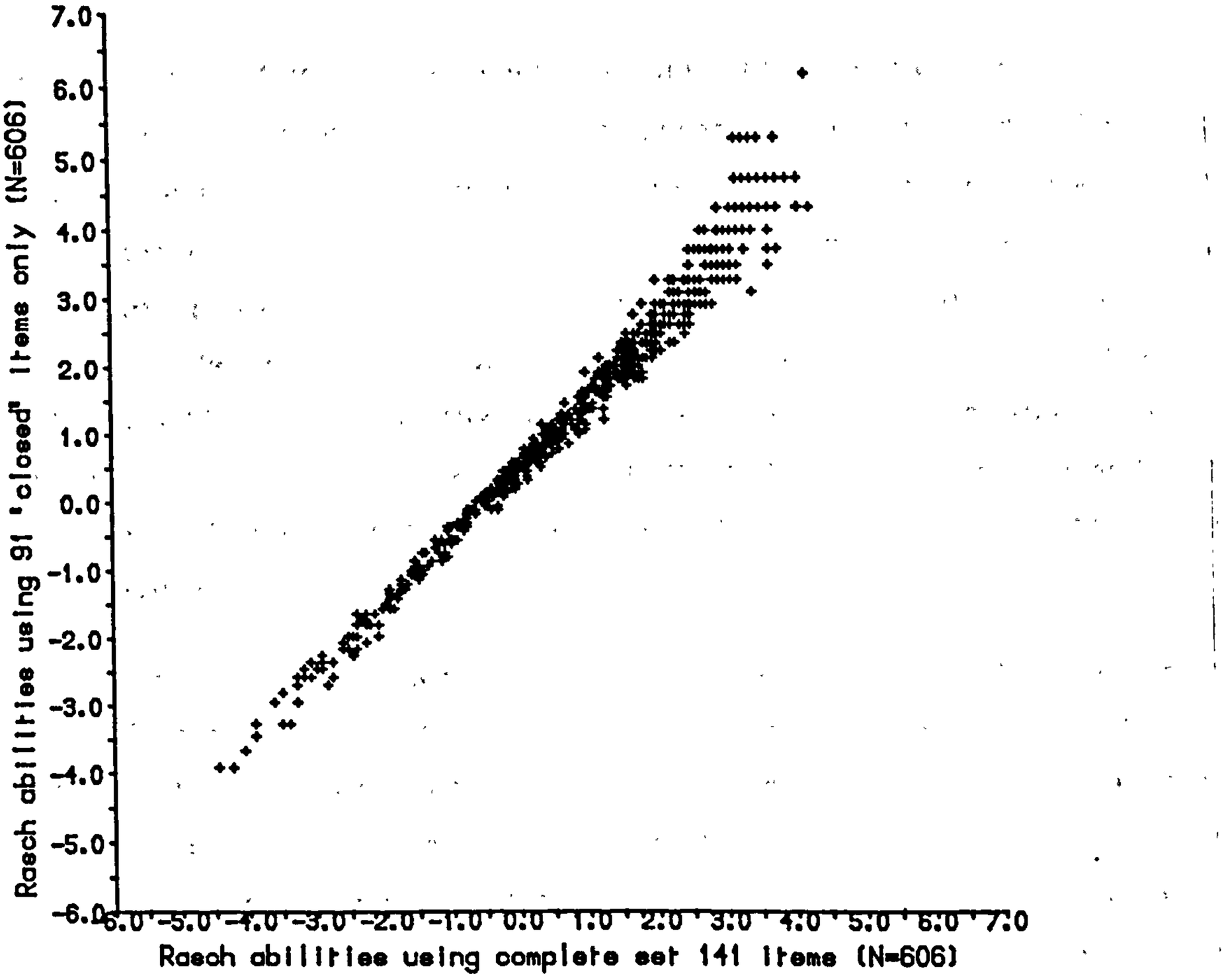


Figure 4.31 Subset- vs Test-Based Abilities Using 'Closed' Items

4.5.2.4 Item Misfit as an Indicator of Departure from Unidimensionality

As was noted in Chapter 3, the examination of items found to show significant misfit can sometimes lead to the discovery of subsets of items with a common feature which sets them apart from the other items in the test. In the work of de Jong (1983), for example, close examination of a number of items found to show poor fit in a test of listening comprehension suggested that these items tapped an ability which might be defined as 'knowledge of the world' or 'general intelligence' rather than listening comprehension.

In the case of the cloze-type test data under discussion here, a number of the instances of item misfit were found to be explicable with reference to inconsistencies or deficiencies in the marking scheme, or to a possible effect of the use of simplified passages with high-level testees (see Section 4.3.2.5). As regards the instances of misfit which could not be accounted for in these ways, it is possible that the items in question required abilities different from those tapped by the test in general; no common feature was discerned in this case, however.

Wright and Stone (1979) point out that the items in a given test may show better fit if analysed in separate subsets, each measuring a different variable. It

follows from this that if one had some rational basis for dividing the complete item set into two or more subsets (as is required in the checks described in the two previous sections), then comparison of the item fit statistics obtained from the subset analyses with those obtained from the whole-test analyses would provide a further way of investigating dimensionality. If items which showed misfit in the whole-test analysis were found to fit well when analysed in a given subset, this would suggest that the subset defined a separate variable. If, on the other hand, the same items showed misfit both in the test as a whole and in the subset, this would indicate that the particular subset isolated did not represent a separate dimension.

Comparisons of this kind were carried out here for the complete item set vs each of the four item subsets created by the content/structure and 'open'/'closed' divisions. All 12 items with total fit t-values of greater than 2 in the content word subset analysis were found to be among the 26 items with total fit t-values of greater than 2 in the whole-test analysis, as were all 12 items showing this degree of misfit in the structure word subset analysis. The same applied to all eight misfitting items in the 'open' item subset, and to 15 of the 17 misfitting items in the 'closed' item subset; the t-values for the two items which did not appear among the 26 least well-fitting items in the whole-test analysis (B14 and K121) were found, however, to be only just above the fit limit of 2 in the subset analysis (2.06 and 2.07 respectively) and only just below it in the whole-test analysis (1.97 and 1.95 respectively). Again, then, it is shown that scores on the item subsets considered here do not represent measures of abilities uncorrelated with those represented by scores on the complete test.

4.5.3 Sample-Independence of Difficulty Estimates

A frequently-mentioned advantage of the use of Rasch analysis is that when the fit between model and data is sufficiently good, the item difficulty estimates will be independent of the particular subpopulation used in calculating them. As was made clear in Chapter 2, this does not mean that difficulty estimates obtained for the same items using different person samples will be exactly the same; it does mean, however, that they should exhibit a certain degree of stability, taking into account the size of the associated standard errors.

Checks for the sample-independence of difficulty estimates for the items in the cloze-type test were carried out here using the method described by Wright and Stone (1979:94-95), in which pairs of difficulty estimates obtained using two different person samples are plotted, and their standard errors used in the

calculation of confidence boundaries. The difficulty estimates compared were obtained separately for testees grouped in two different ways: (i) according to their score on the test, and (ii) by nationality. Following the suggestion of Hambleton and Murray (1983:76), 'baseline' plots, using randomly selected subgroups equivalent in size to those used in the rationally-based groupings, are presented for purposes of comparison.

4.5.3.1 Difficulty Estimates from High- vs Low-Scoring Subgroups

The two subgroups for whom the item difficulty estimates are compared in this section were the 200 highest-scoring and the 200 lowest-scoring persons in the Malaysian data set, i.e. the same subgroups for whom traditional and Rasch indices of item difficulty were compared in Section 4.4.2.2. The plotted difficulty estimates (listed in Appendix G.1) are thus the same as those in Figure 4.19, but with the addition here of 95% confidence boundaries constructed on either side of the identity line by fitting a third order polynomial to the co-ordinate points calculated using Wright and Stone's (1979) method [5]. No more than 5% of the plotted points (in this case 6 or 7 items) should fall outside these limits if sample-independent difficulty estimation has been achieved.

It can be seen from Figure 4.32 below that considerably more points than this fall outside the 95% confidence boundaries, indicating that model-data fit was not in this case sufficiently good for the two sets of difficulty estimates to be considered statistically equivalent. In view of the fact that items identified previously as showing serious misfit have been retained in these analyses, this is not altogether surprising. Furthermore, as was indicated in Section 4.4.2.2, the subgroups used here were somewhat extreme in the areas of the ability scale which they represented, and thus for each subgroup there will have been a rather large number of items whose difficulties could not be accurately estimated, on the grounds of lack of information.

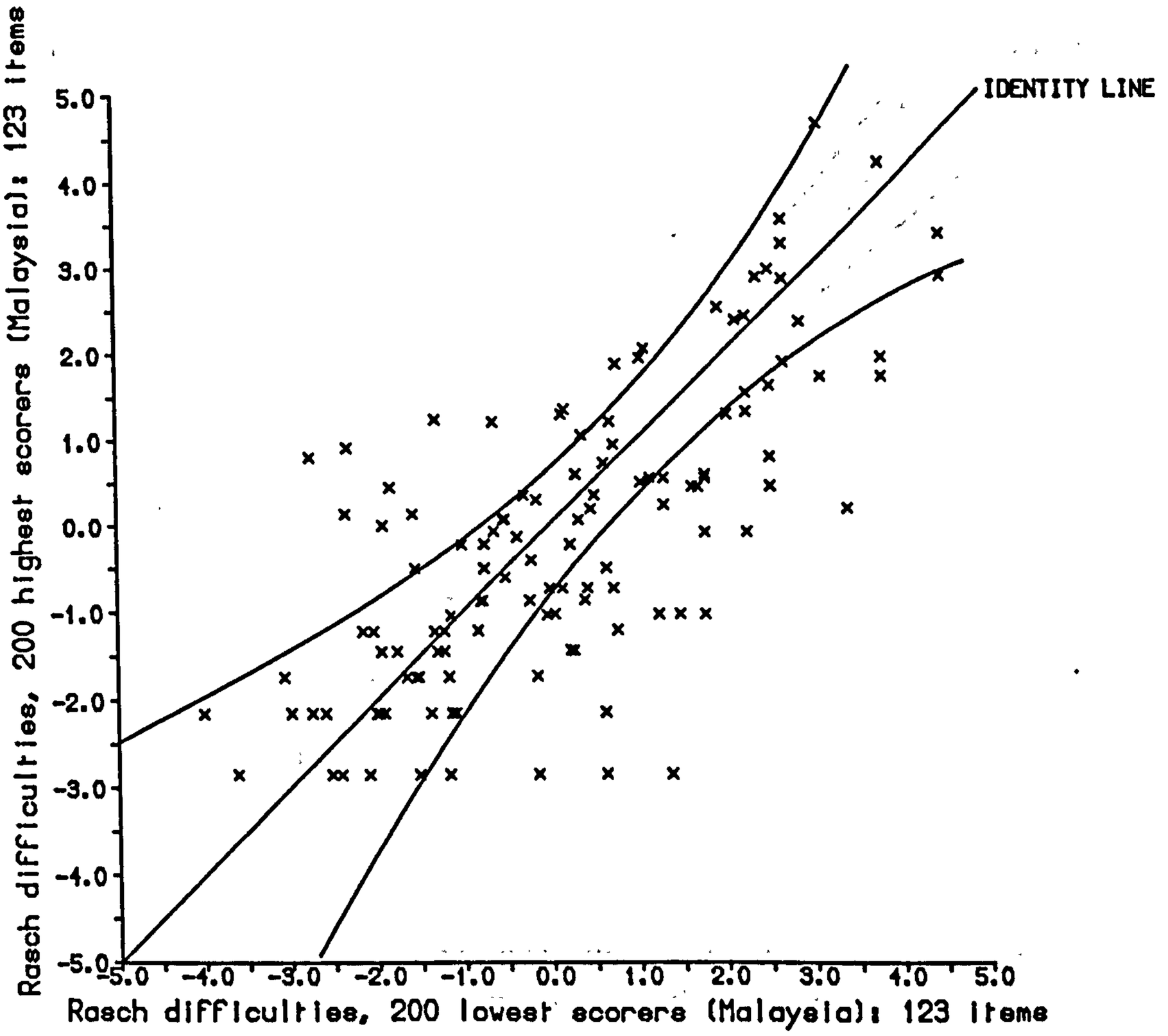


Figure 4.32 Sample-Independence Check, High- vs Low-Scoring Subgroups

In order to demonstrate the effect of having used these particular subgroups for the check described above, the same procedure was carried out using difficulty estimates calculated using two groups of 200 testees drawn at random from the Malaysian sample. The results are shown below, in Figure 4.33.

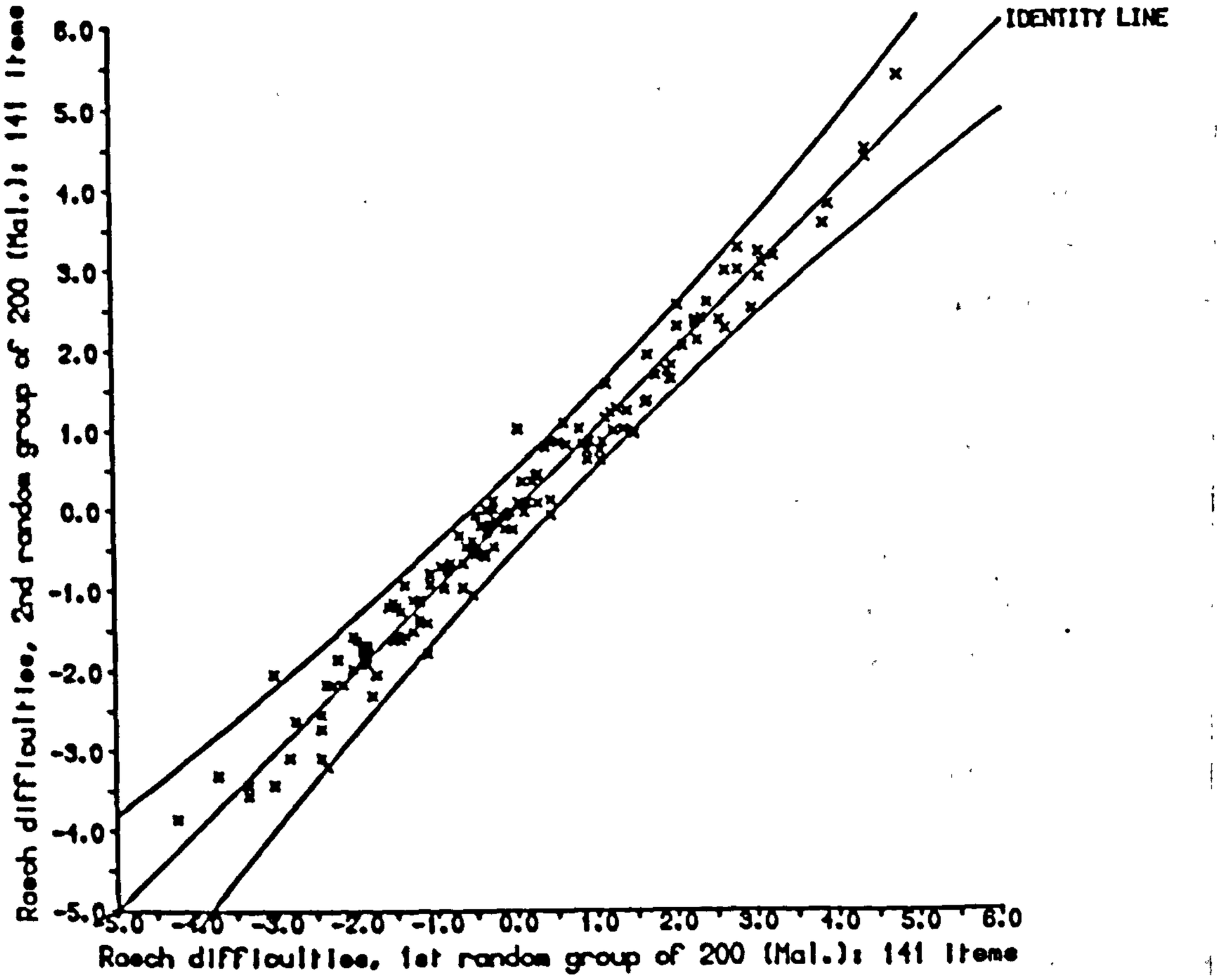


Figure 4.33 'Baseline' Plot Using Random Subgroups (Malaysian Sample)

It can be seen that in this case, very few items fall outside the 95% confidence boundaries, indicating that the estimates obtained using these subgroups are statistically equivalent. (It will be noted that using the random groupings, it was possible to plot points for all 141 items, there being none which were answered all correctly or all incorrectly by either subgroup). The fact that the standard errors of the difficulty estimates were considerably lower in the calibrations based on random subgroups can be seen by comparing the positions of the confidence boundaries in Figures 4.32 and 4.33: since the members of each random subgroup came from throughout the ability range, person abilities and item difficulties were better matched than in the analyses using high- and low-scoring subgroups, with the result that difficulties could be estimated with greater confidence.

4.5.3.2 Difficulty Estimates from Score-Matched Malaysian and Tanzanian Groups

For all previous comparisons of results from the Malaysian and Tanzanian testee groups, the complete samples were used. For the sample-independence check carried out in this section, however, the two data sets were edited so that each contained similar numbers of persons at each observed raw score level. This was done in order to remove the difference between the two original samples in terms of the distribution of abilities, so that the comparison here would be between groups of the same ability levels, but of different cultural, educational and linguistic backgrounds.

These groups, referred to here as 'score-matched' groups, were created by selecting persons of the appropriate score levels at random from the larger data set, to match as closely as possible the levels found in the smaller data set. A few persons also had to be removed from the smaller data set in cases where insufficient persons in the larger set had obtained roughly equivalent scores. The outcome of this editing process was two groups of 234 persons, each with raw scores ranging from 1 to 129. The mean scores for these groups were 61.27 for the Malaysians and 61.30 for the Tanzanians, with standard deviations of 30.93 and 31.07 respectively.

Difficulty estimates for the 141 items were obtained using each of these score-matched samples (see Appendix G.4). These were plotted, together with 95% confidence limits, as before. The results can be seen in Figure 4.34 below; the further slight reductions in the numbers have resulted from the removal of misfitting persons in each case.

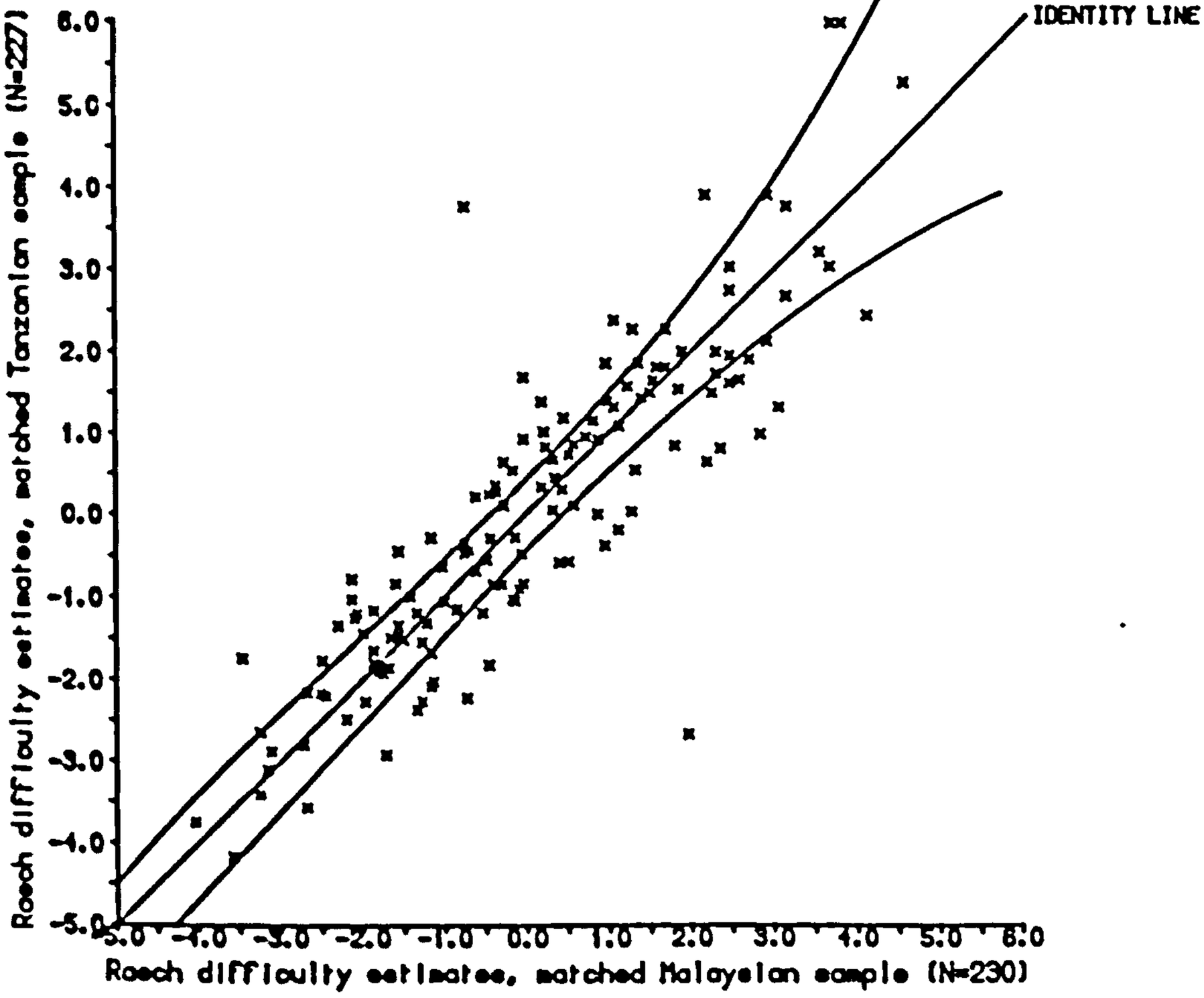


Figure 4.34 Sample-Independence Check, Score-Matched Nationality Groups

As is indicated by the number of points falling outside the confidence boundaries, there is again greater variation in the pairs of estimates than expected, taking into account the standard errors in each case. This will again have been contributed to by the inclusion of items known to show poor fit; the grouping of persons by nationality also appears to have had some effect, however. In Figure 4.35 below, difficulty estimates are plotted for two groups of testees drawn at random from the combined Malaysian and Tanzanian score-matched groups. (The reduction in numbers from 234 to 228 in each case has resulted from the removal of misfitting persons).

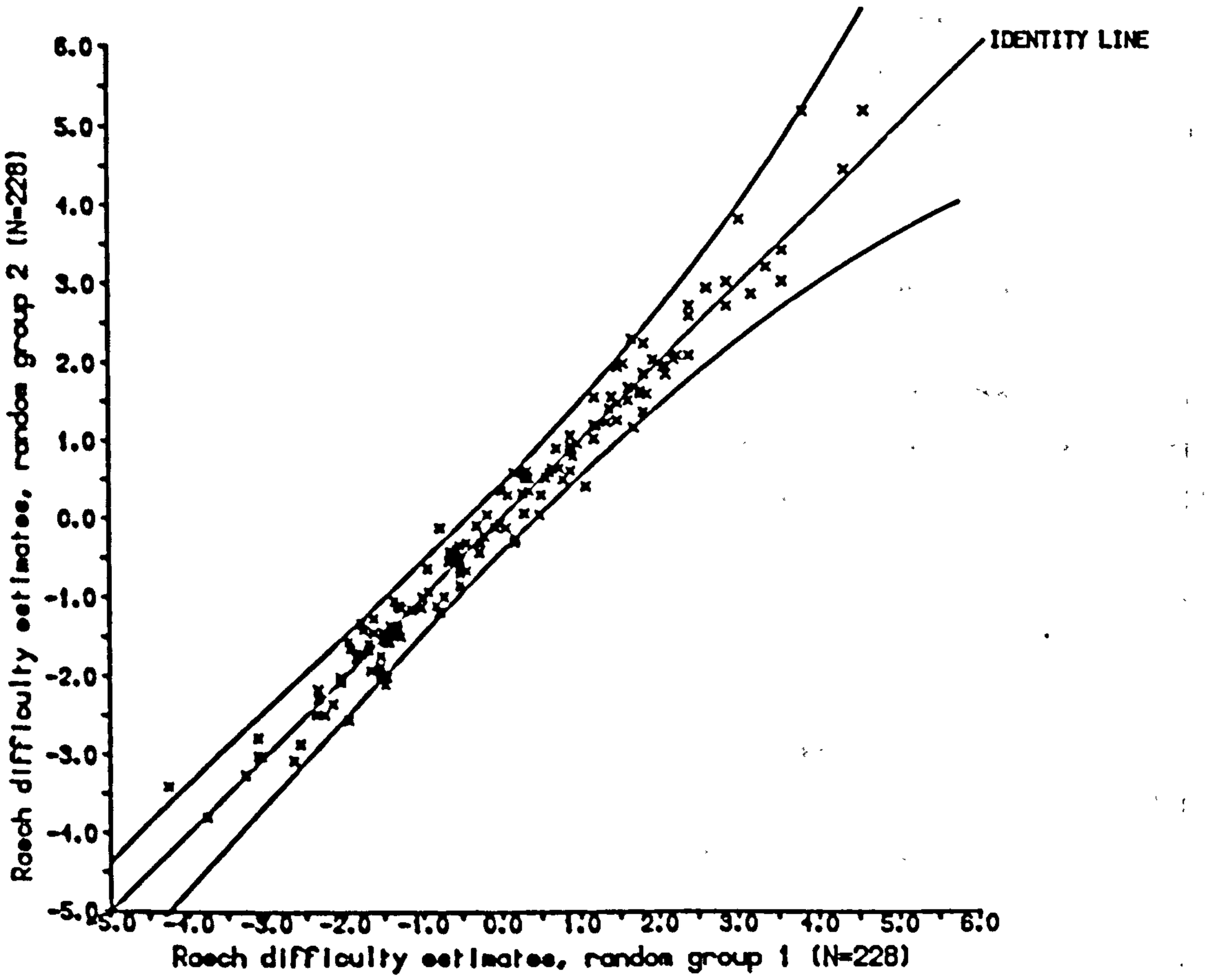


Figure 4.35 'Baseline' Plot Using Random Halves Combined Nationality Groups

The results shown in Figure 4.35 are very similar to those obtained for the 'baseline' plot in the previous section: almost all of the points lie within the calculated limits, indicating that when the total data set is divided in a random fashion rather than by testee background, the obtained estimates can be considered invariant.

From the results presented here, then, it would appear that the sets of difficulty estimates obtained for the two groups at opposite ends of the ability range, and for the two nationality groups, cannot be viewed as sample-independent. Again, it would be of interest to repeat these checks after re-coding the data in accordance with an amended marking scheme, and removing the effects of discrepancies such as that noted in connection with items E59 and L130 (see note 4 at the end of this chapter).

4.5.4 Test-Independence of Ability Estimates

The final investigations of the cloze-type data reported here concerned the stability of ability estimates calculated for the same persons using different subsets of items drawn from the test as a whole. The checks carried out here take comparisons of the kind described in Section 4.5.2.3 a step further in that they involve subsets containing no items in common.

4.5.4.1 Ability Estimates from Content vs Structure Word and 'Open' vs 'Closed' Item Subsets

The ability estimates obtained for the Malaysian testees using the content and structure word item subsets are plotted against each other in Figure 4.36. Points have been plotted here for the 607 persons remaining after the removal of persons scoring zero or full marks on one or both subsets. Figure 4.37 shows the equivalent plot for the 'open' vs 'closed' item subsets; in this case, points have been plotted for 606 persons.

Since the ability estimates shown here have not been adjusted to take into account the difference in mean difficulties of the two items subsets in each case, the identity line is not included in these graphs. The clustering of the points along a straight line is in itself sufficient indication that the pairs of estimates correspond quite closely in both comparisons. The wider spacing of points at both extremes of the ability range, for which reasons were suggested in Section 4.5.2.3, is even more noticeable here, as a result of both subsets in each case being somewhat restricted in the levels of ability which they could differentiate. Nearer the centre of the scale, however, the pairs of estimates can be seen to cluster very closely together, reflecting a high degree of stability.

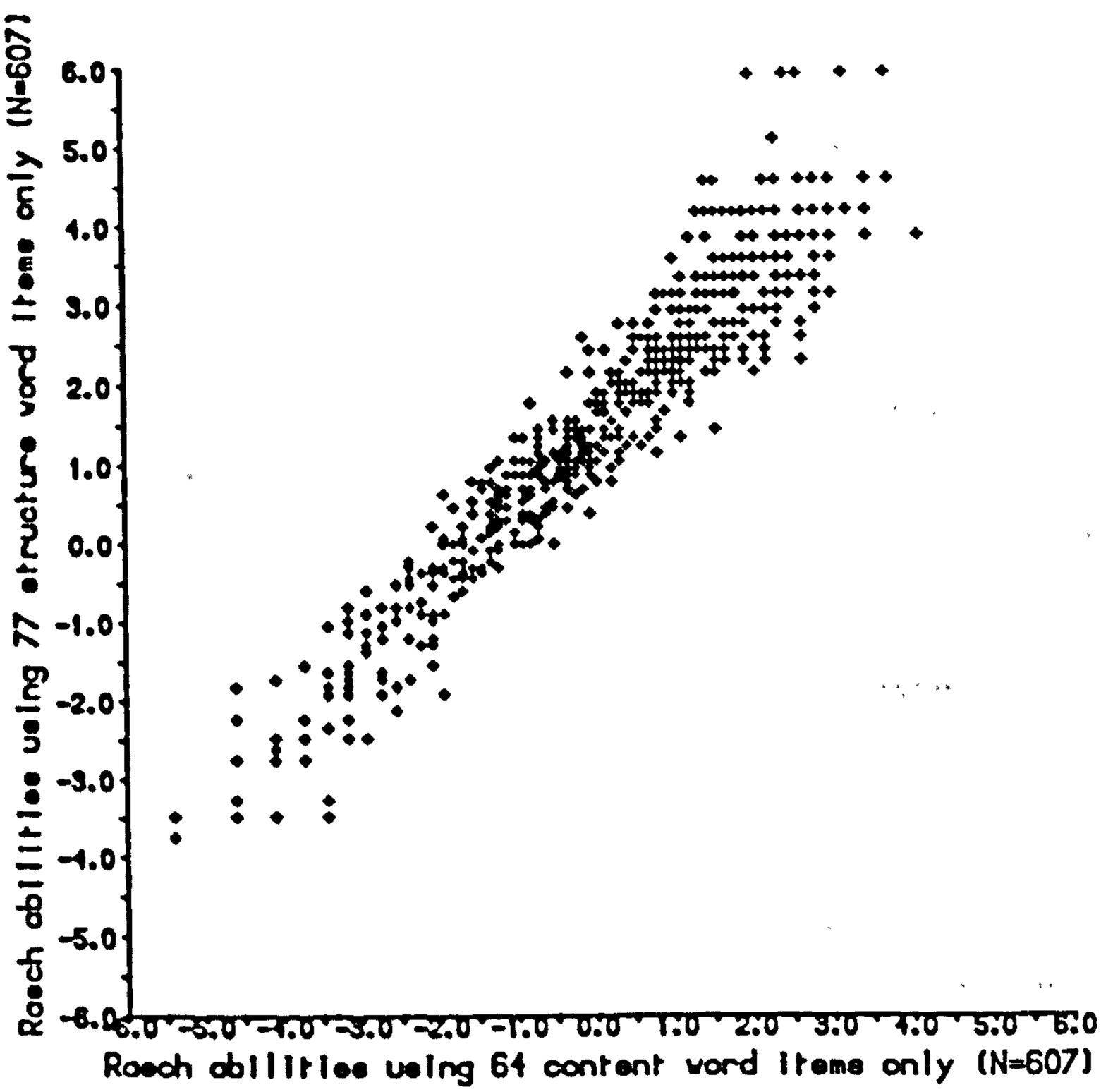


Figure 4.36 Ability Estimates Using Content vs Structure Word Item Subsets

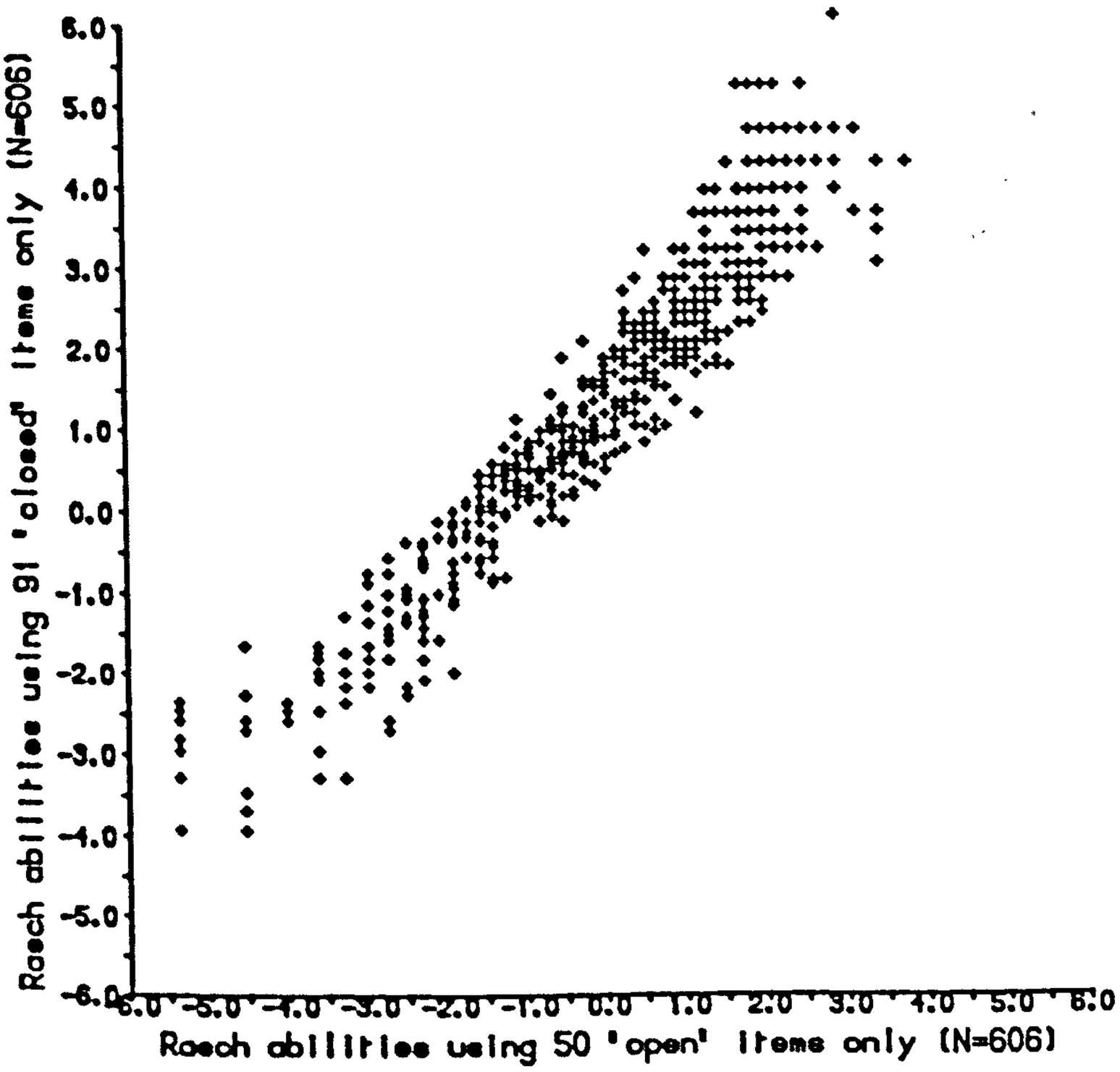


Figure 4.37 Ability Estimates Using 'Open' vs 'Closed' Item Subsets

4.5.4.2 Ability Estimates from Hard vs Easy Item Subsets

For the final analyses reported here, the 40 hardest and the 40 easiest items identified in the whole-test analysis of the Malaysian data were treated as separate tests, and used to estimate the ability of the Malaysian testees. Since these item subsets represented a rather extreme division of the data, a fairly large number of persons had to be excluded from the comparison on the grounds of having scored zero or full marks on one or other of the subsets. After the removal of these, 425 persons remained.

For a comparison of this kind one would, ideally, estimate the abilities using the difficulty estimates obtained for the two sets of items in the whole-test calibration. Since this is not possible using BICAL, the two sets of ability estimates were obtained from separate calibrations, and subsequently adjusted to take account of the differences in the difficulties of the item subsets when estimated in the whole-test calibration and when calibrated separately. In each of the separate calibrations, the values from which are listed in Appendix G.5, the mean difficulty was, as usual, set to zero. In the whole-test calibration, on the other hand, the mean difficulties for the hard and easy subsets were 2.29 and -2.09 respectively. When compared individually, it was found that the differences between the two estimates for each item could, taking into account the standard errors, be considered equivalent to the difference between means for the separate calibrations. It was, therefore, possible to adjust each ability estimate by adding, for the relevant item subset, the mean difficulty obtained from the whole-test calibration.

The adjusted ability estimates from the hard and easy item subsets are shown in Figure 4.38 below. It can be seen that, since this is such an extreme comparison, few persons have been satisfactorily measured by both 'tests'. In most cases, either one subset or the other has been so ill-matched with the persons' abilities as to make differentiation of levels impossible. It is for this reason that the plotted points form the horizontal and vertical lines seen in Figure 4.38.

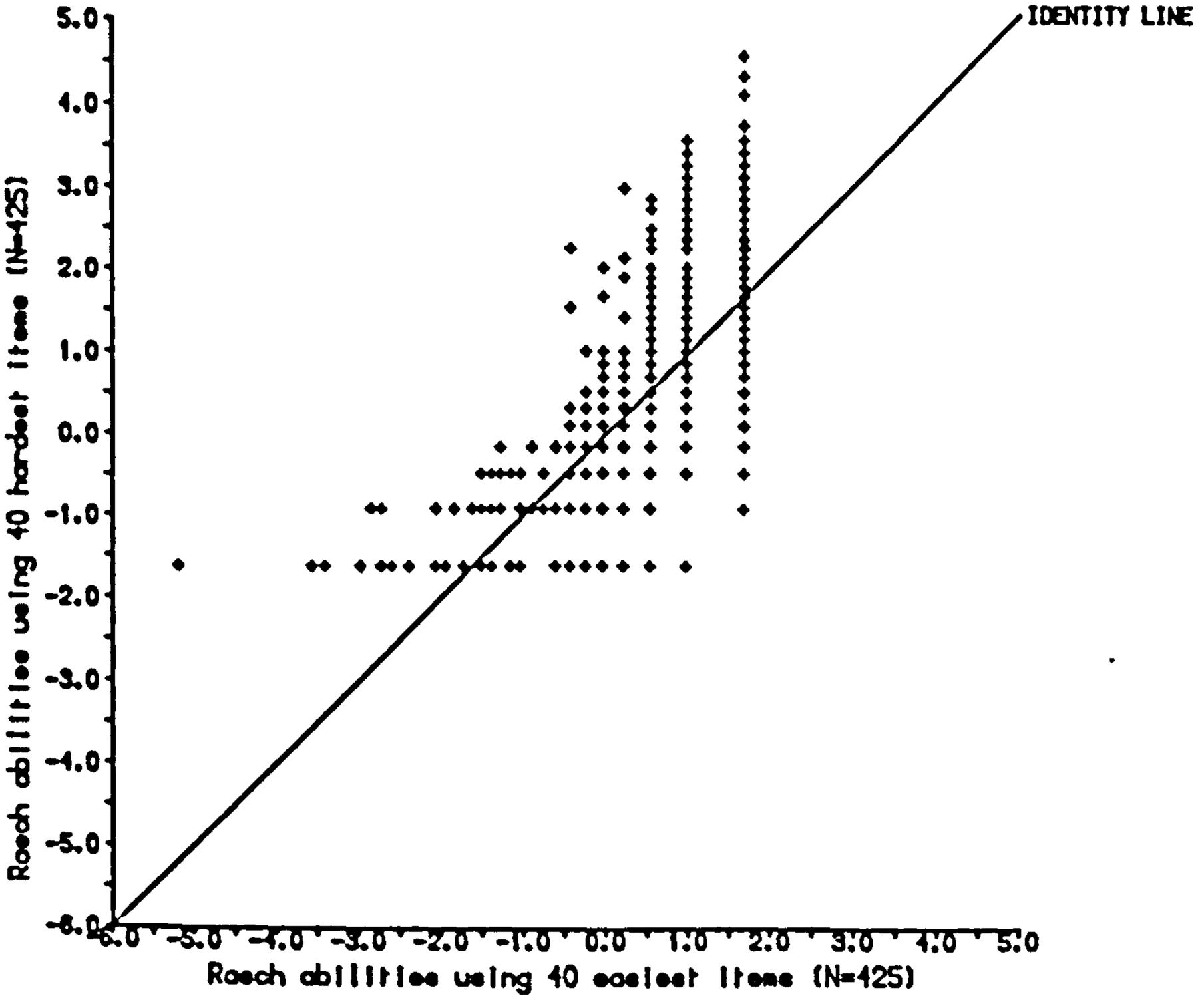


Figure 4.38 Ability Estimates Using Hard vs Easy Item Subsets

The persons for whom the points fall in the central portion of this graph are the only ones for whom reasonable measures have been obtained from both item subsets; it is thus only in these cases that one might speak of the test-independence of the ability estimates. As a check on the consistency of ability measures across different item sets, then, this example is not informative; it does, however, underline the need for item sets to be of suitable difficulty levels for the persons concerned, in order for appropriate measures to be possible.

Given the wide range of abilities represented in this data set, one might expect it to be difficult to find a subset of only 40 items which could yield appropriate measures for most persons. For purposes of comparison with the hard vs easy subsets described above, two random sets of 40 items were selected from the complete set, and ability estimates obtained using each of these. It was possible to retain 602 persons in these analyses.

In this case, the differences between the difficulties from the whole-test calibrations and the subset calibrations of the item sets were negligible; the appropriate adjustments were, however, made to the ability estimates for the sake of consistency with the previous example. The results are shown in Figure 4.39 below.

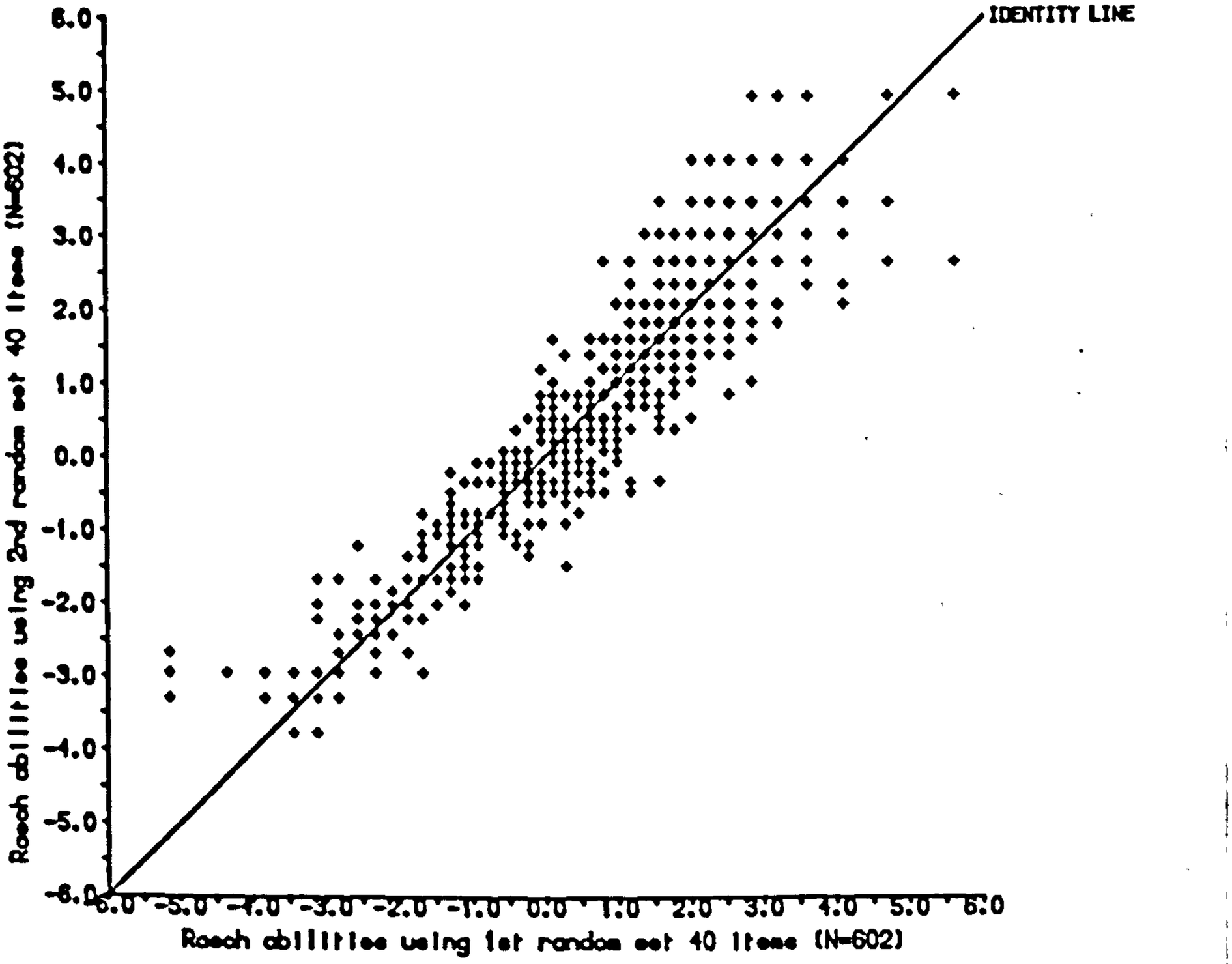


Figure 4.39 Ability Estimates Using Random Item Subsets

Although there is a portion of the graph in which the points cluster about the identity line, the pattern of horizontal and vertical lines referred to in previous examples is again evident, though, of course, less so than in the hard vs easy item comparison. The graph indicates that, particularly towards the upper end of the ability range, both item sets have been somewhat restricted in their capacity to differentiate among levels. This can be attributed, at least in part, to their restricted size: in each case, the 602 persons could be placed at only 39 different points on the ability scale, which, in view of the observed distribution of scores for the Malaysian sample on the test as a whole, means that only rough divisions could be made. Again, then, a less stringent check, perhaps using halves of the

complete item set, would have been more informative in terms of the independence of ability estimates from particular item subsets.

4.6 Summary of Findings

In discussing traditional and Rasch analyses of the cloze-type test data, this chapter has sought firstly to provide a comparison of the two approaches in terms of the nature and usefulness of the information they yield, and secondly to consider the extent to which the advantages of Rasch analysis, obtainable in theory, have been realised in this particular application.

It was shown that while the results of the Rasch analysis were not inconsistent with those of the traditional analysis, there were several ways in which the Rasch approach proved more informative. The item fit statistics, for example, provided a clearer indication of which items gave rise to odd response patterns, since, unlike the traditional indices of discrimination, they were not bound up with item difficulty. Information of the kind obtained fom the analysis of person fit was not available at all under the traditional approach; in view of its implications in terms of the validity of test scores, however, such information is clearly of value. Other points in favour of the Rasch approach included the more rational treatment of measurement error and the usefulness of placing abilities and difficulties on the same scale.

Their greater stability across different testee groups was noted as a further advantage offered by the Rasch difficulty estimates over the traditional indices of item difficulty. It was, however, found that the sample-independence of difficulty estimates which would allow these items (or passages) to be characterised e.g. for purposes of item banking was not wholly achieved for the data sets analysed here. This was attributed in part to the mis-match between persons and items in the comparisons involving person subgroups and item subsets from opposite extremes of the ability/difficulty range, and in part to lack of fit between model and data. Although many of the items showed reasonable fit, some were found to deviate from expectation, in some cases quite markedly. On the basis of the apparent behaviour of these, a number of possible improvements to the test were proposed.

No evidence was found to suggest that the cloze-type data violated the assumption of unidimensionality. The assumption of local independence was not investigated specifically; one way in which this might be checked, however, would be by comparing the results of analyses of the items treated separately,

and of the same items treated as clusters within their passages (in the manner described by Andrich et al., 1982).

The use of the partial credit version of the Rasch model is suggested by Pollitt and Hutchinson (1987:91) as a means of controlling for the possible interdependence of cloze items. They further note the potential application of partial credit analysis in the scoring of cloze items. In examining some of the responses for the purposes of this study it was indeed evident that information was lost in scoring these items dichotomously: had the scoring method given partial credit for answers which were only either syntactically or semantically acceptable, the difference in the nature of the 'wrong' answers given e.g. by those who had not understood the passages and those who had attempted over-sophisticated answers would have been reflected. Although a more refined scoring method might not be necessary, or even practicable, for this particular test, it was undoubtedly the case that within the broad category of 'wrong' answers, additional information about proficiency was available.

A further observation made in connection with the scoring of this test was that in cases where the marking scheme gave no credit for acceptable answers (and thus operated rather as though exact word marking had been used), it was usually the higher-level testees who were penalised. It seems likely, therefore, that use of acceptable word scoring for cloze tests will result in better fit to the Rasch model than use of the exact word scoring method.

Notes on Chapter 4

1. Only certain parts of the BICAL output have been selected for discussion here; this list of Rasch statistics does not, therefore, represent the complete BICAL output.
2. The magnitude of the standardized residual is given as 9 in the BICAL output even where the actual value is larger than this (see Wright et al., 1980:80): values of this magnitude will always be noteworthy, and so beyond a certain point the precise value becomes almost immaterial. The value of 9 referred to here, however, is in fact the actual one.
3. Preliminary versions of these results, and of some of those reported in Section 4.5, appeared in Woods & Baker (1985), a copy of which is appended (see Appendix K).
4. The low facility value for the Malaysian sample on item E59 ("Yes, but look at (E59) rock!") resulted from the fact that one of the answers on the

marking sheet ('the') had not been counted as correct. Since over half the sample had given this answer, the proportion correct should actually have been approximately .9, and not .3. For the Tanzanian sample, the marking scheme had been applied correctly; the misfit shown by this item in the Rasch analysis for this group can be attributed to the fact that the acceptable answer 'this', chosen by a number of the higher-scoring persons, was not included in the marking scheme. (The other extreme outlier, on the opposite side of the identity line, is item L130: 'The Air Hostess went away and came back with a (L130) of whisky.' Only 'glass' appeared on the marking sheet; a very common choice of answer, however, made by both groups, was 'bottle', which was acceptable in the context. It would appear that this was marked correct for the Malaysian group, but not for the Tanzanian group: had the marking been consistent, the facility value for the latter group would have been approximately .5, and not 0.03.)

5. For each pair of difficulty estimates plotted, the 95% confidence limits correspond to points lying on a line drawn through the item point, perpendicular to the identity line, at distances of $1.96 \times [(s_{i1}^2 + s_{i2}^2)/2]^{1/2}$ on either side of it, where s_{i1} and s_{i2} are the standard errors of the two difficulty estimates.

CHAPTER 5
ANALYSIS OF ELTS TEST DATA

In this second set of analyses, similar methods of analysis are applied to data from a different type of English proficiency test. In some respects, therefore, this chapter runs parallel to the previous one. However, for reasons of test security it is not possible to append a copy of the ELTS test, or to consider the content of particular items. Discussion of the results of the traditional and Rasch analyses is therefore necessarily briefer than in the previous chapter, and somewhat different in nature: since this part of the study involved 8 separate subtests, these results are, where possible, presented in summary form, and explanatory detail already provided in Chapter 4 is kept to a minimum.

5.1 Description of the ELTS Data

The data for the analyses presented in this chapter were obtained from Pattern A of the English Language Testing Service (ELTS) test, which was developed jointly by the British Council and the University of Cambridge Local Examinations Syndicate. The purpose of this test, which has been in operation since 1980, is to assess the level of proficiency of non-native speakers of English wishing to enter higher education in Britain.

5.1.1 Composition of the Test

The overall structure of the test, as set out in *English Language Testing Service: An Introduction*, is as follows:

1. A general (G) section, intended to test general proficiency in English, and consisting of two subtests: G1 (Reading) and G2 (Listening).
2. A modular (M) section, intended to test study-related language skills, and consisting of three subtests: M1 ('Study Skills'), M2 (Writing) and M3 (Interview).
3. The modular section taken by each candidate is chosen, according to his/her subject area, from the following:

General Academic	(GA)
Life Sciences	(LS)
Medicine	(ME)
Physical Sciences	(PS)
Social Studies	(SS)
Technology	(TN)

Since this study is concerned only with dichotomously-scored items, subtests G1, G2 and M1, which are composed entirely of 4-option multiple-choice items, were the only parts used. (M2 and M3 are scored on rating scales, and are therefore not suitable for the kind of item analysis presented here.)

G1 (Reading) contains 40 items, and consists of the following subsections:

1. Items 1-11: Choosing the most accurate paraphrases of given sentences;
2. Items 12-24: Choosing the correct words to complete single-word gaps in a passage;
3. Items 25-40: Answering reading comprehension questions on 3 different newspaper reports of the same event.

G2 (Listening) contains 35 items, and is divided into subsections as follows:

1. Items 1-10: Choosing the correct diagrams from taped descriptions;
2. Items 11-16: Answering comprehension questions (presented in written form) on a taped interview;
3. Items 17-26: Choosing appropriate (written) replies to taped questions;
4. Items 27-35: Answering comprehension questions (presented in written form) on a taped 'seminar'.

For M1 (the modular reading subtest), the items relate to passages and diagrams appearing in a booklet of texts drawn from books, journals and reports. There is a separate source booklet for each of the 6 subject areas listed above. The 6 different M1 subtests based on these each consist of 40 comprehension items.

In this discussion, each item is identified by a prefix which denotes the subtest to which it belongs (G1, G2, GA, LS, ME, PS, SS, TN), and a two-digit number indicating its position within that subtest. Thus the items in subtest G1, for example, are referred to as G101, G102 etc.

5.1.2 Administration and Scoring

The time limits set for G1 and M1 are 40 and 55 minutes respectively. G2 takes approximately 30 minutes to administer; the taped material is heard only once, and candidates complete their answer sheets during pauses left on the tape.

Scoring is done manually, using templates. The number-correct score obtained on each subtest is converted to a band score on a 0 - 9 scale corresponding to the scales used for rating the written composition and interview sections of the test. For the purposes of this study, however, these band scores

are not relevant, and only the response patterns and the raw scores are used.

5.1.3 Description of Sample

The responses of 1,503 non-native speakers of English, tested in Britain between 1981 and 1986, formed the data set for this part of the study. The sample was the same as that used in the item analysis sections of the ELTS validation study conducted by Criper and Davies (1986), but with the addition of 178 candidates tested in the academic year 1985-6.

The candidates, who came from a wide variety of different countries and language backgrounds, all took the ELTS test shortly before beginning courses of higher education in Britain. It should be noted that this sample cannot be considered representative of the target population for the test: since these candidates were all tested after their arrival in Britain, and hence after acceptance for study, it is likely that their average proficiency level will be higher than that for the total target population.

All candidates took G1 and G2, plus one of the 6 subject-related M1 subtests. This study is therefore concerned with responses on a total of 8 different subtests, with sample sizes as follows:

Subtest	No. of Testees
G1 (Reading)	1,503
G2 (Listening)	1,503
M1 (General Academic)	403
M1 (Life Sciences)	374
M1 (Medicine)	143
M1 (Physical Sciences)	134
M1 (Social Studies)	264
M1 (Technology)	185

For the analyses carried out here, all responses were coded by number, according to the particular multiple-choice option chosen, with a separate code for omitted answers. It was thus possible to obtain a tally for each of the distractors, as well as categorising each response as correct, incorrect or omitted.

As was the case in the previous chapter, the basic data sets were combined and subdivided in various ways for the purposes of particular analyses. Details of the new sets thus formed are, as before, given in the relevant sections below.

5.2 Traditional Analysis of ELTS Data

5.2.1 Traditional Statistics Computed

For each of the 8 subtests listed above, traditional statistics were computed as in Chapter 4:

1. Raw score for each person,
2. Facility value for each item,
3. E_{1-3} discrimination index for each item,
4. Unbiased point biserial correlation coefficient for each item,
5. K-R20 internal consistency reliability coefficient,
6. Standard error of measurement.

5.2.2 Summary and Interpretation of Results

The results of these analyses are set out in Appendices H.1 to H.3. As before, the person scores are summarised in the form of frequency counts and histograms, and the item statistics listed both individually and grouped by interval.

5.2.2.1 Raw Score Distribution

The raw score distributions for the 8 subtests are shown in Tables 1-8 in Appendix H.1. It can be seen from Tables 1 and 2 that the distributions for the two General subtests are similar, with relatively few persons at the lower end of the raw score scales, and the highest concentration in the mid- to upper portions. (Since the items are all multiple-choice, and hence allow chance success from random guessing, it is to be expected that scores at the low extreme will be rare.) Scores were, however, generally higher on G1 (Reading) than on G2 (Listening): scores on the latter extended lower in the possible range, and while 23 of the 1,503 candidates scored 40 on G1, no candidate obtained the maximum score on G2. The mean scores for G1 and G2, expressed as percentages (since these two subtests are of different lengths), were 70.5% and 67.1% respectively.

The score distributions for the 6 Modular subtests (see Tables 3-8 in Appendix H.1) are not directly comparable, since they concern different groups of persons who cannot be considered representative of the populations taking the various modules. However, in order to gain an impression of the nature of the data, it is of interest to note the main features of each.

The score distributions for the General Academic and the Social Studies modules have the shape almost of normal distribution curves, though with fewer persons near the low extreme than near the high extreme, and with particularly high frequency counts for some scores. The mean scores (21.2 and 21.9 respectively) are in both cases very close to the midpoint of the available range. The score distributions for the Medicine, Physical Sciences and Technology modules, on the other hand, are all negatively skewed; this is particularly noticeable in the case of the Technology module, for which the scores are less evenly distributed in the upper half of the scale than for the Medicine and Physical Sciences modules. In all three cases, the mean scores are well above the midpoint of the scale, ranging from 28.5 to 29.8. The distribution for the Life Sciences module falls somewhere between the two basic patterns identified so far. It is similar in shape to those for the General Academic and Social Studies modules, but is again negatively skewed, though less so than the other 3 distributions; this is reflected in the somewhat lower mean score (24.6) observed for the Life Sciences module.

It is not, of course, possible to say whether these differences arise from differences in the difficulty of the 6 Modular subtests, or from differences in the proficiency levels of the person subgroups. Although it would be possible to investigate this by considering the same subgroups' responses on the two General subtests, this is not the purpose of the present study.

5.2.2.2 Item Facility Values

The facility values for all items in the 8 ELTS subtests are listed in Tables 1-8 in Appendix H.2. The general distributions of these can be seen from Tables 1(a) to 8(a) in Appendix H.3.

These tables show that for both General subtests, all but a few items are placed at or above the midpoint of the facility value scale, and that in both cases over half the items have facility values of .7 or above. These subtests thus proved relatively easy for this sample. As was indicated in the previous section, G1 appears in general slightly easier than G2: no item in G1 has a facility value of less than .4, while those falling below the midpoint in G2 have values ranging from .1 to .41.

As regards the facility values of the items in the 6 Modular subtests, the distributions of these are, of course, largely predictable from the score distributions referred to above. Only the General Academic, Social Studies and

Life Sciences subtests have roughly equal numbers of items on either side of the midpoint of the facility scale; for the other 3 subtests the values are almost all at or above .5, with over half the values in each case being as high as .7 or above. The difference in the difficulty of these subtests, or in the proficiency of the subgroups, is thus again apparent.

5.2.2.3 Indices of Discrimination

The E_{1-3} indices of discrimination and the unbiased point biserial correlation coefficients for each item are listed in Tables 1-8 in Appendix H.2, and summarised in Tables 1(b) to 8(b) and Tables 1(c) to 8(c) in Appendix H.3.

Comparison of Table (b) with Table (c) in each case shows that the extent to which the same items are identified as poor discriminators by both indices varies from one subtest to another. In the case of G1, for example, the 10 items with the lowest values on each index contain only 5 items in common, and the poorest discriminator identified by each differs. For M1(GA), on the other hand, 9 of the same items appear among the 10 poorest discriminators identified by each index, with the lowest value on each being associated with the same item.

For the sake of clarity, this discussion of discrimination will be based on only one of these indices. Since the point biserial index was found in the analyses reported in Chapter 4 to be less affected by the characteristics of particular groups than the E_{1-3} index, and since it is more widely used in practice, it is the point biserial which will be referred to here.

Comparison of Tables 1(c) and 2(c) in Appendix H.3 indicates that the items in G1 show generally slightly higher discrimination than those in G2; indeed, for G2 the highest point biserial observed is only .39, and one item in particular (G227) is identified as showing reverse discrimination (point biserial = -0.15). As regards the relationship between low discrimination and item difficulty for these two subtests, examination of the associated facility values reveals that the only item in G1 with a point biserial below .2 (G111) was neither of extreme easiness nor extreme difficulty (f.v. = .41). The 9 items with point biserials from .2 to .29, however, were mostly among the easiest items in this subtest; only 2 of these (G109 and G125) have facility values of less than .8. Thus of the 10 lowest discriminators in G1, 3 appear to warrant particular investigation. Of the 3 items with point biserials below .2 in G2, one (G227) is of extreme difficulty (f.v. = .1), and one is among the easiest items in the subtest, though not quite at the extreme (G226; f.v. = .83). The third, item G235, with a facility value of .3, would

appear to be failing to discriminate for reasons other than its difficulty, and is therefore questionable. Item G227 also requires closer inspection, despite its extreme difficulty, since the negative point biserial indicates that the (relatively) few persons who answered it correctly were not in general among the highest-scoring persons.

Negative point biserials were observed also for one item in each of 3 of the M1 subtests (GA, LS and TN). Only in one case (LS35) was this coupled with an extremely low facility value (.05); in the other two cases (GA03 and TN13) the facility values were .37 and .48 respectively. Indeed, it is found that of the 36 items with point biserials of less than .2 in the M1 subtests, only 9 have facility values of less than .2 or greater than .8. Although the same limit for the point biserial would not necessarily be appropriate for all 6 subtests (because of differences in the degree of homogeneity of the various item sets and person samples), it nevertheless appears from these results that each M1 subtest contains at least a few items which do not show consistency with their respective item sets.

5.2.2.4 Test Reliability and Error of Measurement

The K-R20 coefficients of internal consistency reliability and the standard errors of measurement for each of the 8 subtests are shown at the foot of Tables 1-8 in Appendix H.1. The K-R20 coefficients range from .80 (for G2) to .90 (for M1(PS)), and the standard errors of measurement from 2.3 (for M1(PS)) to 2.8 (for M1(SS)).

As was explained in Chapter 4 (Section 4.2.2.4), the K-R20 is influenced by a number of factors relating to the person sample and item set; it is likely that the lower values observed in the ELTS analyses result from the greater homogeneity of the person samples and the smaller size of the item sets. The standard errors of measurement for the ELTS data sets are correspondingly higher than those for the cloze-type test data sets, taking into account the difference in length between the ELTS subtests and the cloze-type test.

5.3 Rasch Analysis of ELTS Data

5.3.1 Rasch Statistics Computed

For each of the 8 ELTS subtests, Rasch statistics were calculated as in Chapter 4:

1. Ability estimate and standard error for each raw score,
2. Weighted total fit t-statistic for each person,
3. Difficulty estimate and standard error for each item,
4. Observed ICCs across 6 raw-score groups, and proportional departures from expectation,
5. Weighted total fit t-statistic for each item,
6. Between-group fit t-statistic for each item,
7. Person separability index/Test reliability of person separation,
8. Number of person strata.

5.3.2 Summary and Interpretation of Results

The results of the Rasch analyses of the ELTS subtests are set out in Appendix I. Where sample sizes are slightly different from those referred to earlier, this results from the exclusion of persons scoring zero or full marks on a given subtest. Numbers of persons with zero scores were 1 on M1(LS) and 1 on M1(PS), while numbers with perfect scores were 23 on G1, 1 on M1(GA) and 1 on M1(TN).

5.3.2.1 Person Ability Estimates

Although it is sometimes recommended (see Wright et al., 1980) that for purposes of item analysis, persons with scores which could have been achieved by random guessing be excluded from the data set, no such editing was carried out for this study, since it is of interest here to see whether evidence of guessing can be found in the results of the analyses.

The raw score-to-ability conversion tables for each subtest are given in Appendix I.1. Two tables are shown for each subtest: the first calculated using the responses of all measurable persons, and the second calculated after the removal of misfitting persons. It can be seen that the two sets of estimates in each case differ only minimally; as in Chapter 4, only the second (i.e. final) set will be referred to here.

For each subtest, the Rasch ability scale ranges from approximately -4 to +4

logits. Standard errors of ability range from approximately .35 (for those near the centre of the raw score ranges) to just over 1 (for those at the extremes). The standard errors near the centre of the ability range are larger than those for the cloze-type test because of the smaller numbers of items involved here.

The observed ability ranges for these ELTS candidates do not span the whole of the available range: as can be seen from the information given at the foot of each table, the lowest ability observed in any subtest is -2.38 (for G2). For all subtests, though, observed abilities extend in the opposite direction to at least +3, and in most cases to approximately +4. This simply reflects the fact that scores on these subtests were rarely at the low extreme, but extended in all cases to (or at least near to) the upper extreme.

5.3.2.2 Person Fit

The numbers of persons with weighted total fit t-statistics of greater than 2 are shown for each subtest in Table 5.1 below. (For convenience, the M1 subtests will from now on be referred to only by their subject area names.)

Subtest	No. of Misfitting Persons	No. of Measurable Persons	% of Misfitting Persons
G1	32	1,480	2.2%
G2	22	1,503	1.5%
GA	10	402	2.5%
LS	18	373	4.8%
ME	1	143	0.7%
PS	1	133	0.8%
SS	7	264	2.7%
TN	2	184	1.1%

Table 5.1 Numbers of Misfitting Persons in ELTS Data Sets

The percentages of misfitting persons in the different data sets range from .7% (for ME) to 4.8% (for LS). In view of the relatively large numbers of persons involved, the person statistics and standardized residuals are not listed individually for the misfitting persons identified in these analyses; the main points noted from inspection of these are, however, summarised below.

All of the misfitting persons in the G1 data set have estimated abilities which are below the mean for the group, and almost half fall below the midpoint of the ability scale for the subtest. This might suggest that misfit is due to correct

guessing on the part of relatively low-level persons. However, although there are some possible signs of this in the standardized residuals (in the form of the occasional fairly large positive value), the predominant pattern for all of these persons is one of negative residuals, particularly near the beginning of the test. Since this was the first subtest administered, it is possible that a major reason for the misfit observed here was the initial unfamiliarity of the test procedure. If this is the case, the abilities of these persons will, on the whole, have been somewhat underestimated, with the result that some later successes appear unexpected when in fact they should not.

None of the same candidates showed significant misfit on both G1 and G2, but all but one of those identified as misfitting on G2 were again below the mean in ability, and over half had estimated abilities below the midpoint for the subtest. As in G1, the residuals are, on the whole, either zero or positive at the end of the test; there is, however, stronger evidence of chance success in the residuals for G2, in that some of the positive values are very large. While the largest value observed in the analysis of G1 was 3, the values for G2 include several of 4, 5, 6, 7 and 8. A difference noted in the general patterns of residuals for these two subtests is that for G2, the strings of negative values tend to occur in the middle of the test rather than at the beginning. Indeed, a number of persons have strings of negative values beginning around item G217; this, it is interesting to note, corresponds with the beginning of a new item subset ('Replying to Questions') within the subtest. It is possible that the change of task was in itself sufficient to impair the performance of some persons; an alternative explanation would be that the particular item type caused confusion.

None of the 10 persons identified as misfitting in the analysis of the GA subtest showed misfit in either of the General subtests. Again, though, the majority were of below mean ability. As regards the residuals for these persons, it is again noticeable that positive values occur mostly at the end of the subtest, and as was the case for G1, there are no extreme values in either direction (values range from -3 to +2). Negative values appear to be scattered throughout; there are, however, several persons with a sequence of negative residuals fairly near to the beginning of the subtest. Again, this appears to correspond with the beginning of a new item subset, in this case relating to a different text in the source booklet. While the transition itself may have contributed to misfit here, it is possible also that some feature of the text and/or the item subset exerted an influence. One way in which these items differ from the others in GA is that they require the interpretation of information presented in tabular and graphical

form, a task in which some of the persons taking this module may have had little practice.

The LS subtest has a larger proportion of misfitting persons than any other module, and 4 of the 18 persons identified also showed misfit on either G1 or G2, suggesting, perhaps, an oddity in the general strategy of some persons. Again, almost all of the persons identified were of below mean ability. As was also the case for G2, residuals extend further in the positive direction than in the negative direction (values range from −4 to +8). Thus some of the correct answers were considerably more unexpected than any of the incorrect answers, an indication that chance successes may have occurred.

The person identified as misfitting on the ME subtest had a very low score (9/40), and fell approximately 3 standard deviations below the mean ability for the group. The list of residuals for this person contains no values below −1, but several values of +3 and one of +6, a pattern which suggests that s/he made a number of correct guesses.

The pattern for the misfitting person on the PS subtest is of a different type. This person's estimated ability is less than one standard deviation below the group mean, and the pattern of residuals indicates that s/he made a larger number of unexpected incorrect answers than unexpected correct ones.

Of the 7 misfitting persons identified in the analysis of the SS subtest, 1 also showed misfit on G1, and almost all had estimated abilities below the mean. As in some of the cases already described, there is some evidence of chance success towards the end of the test, and the residuals which are largest in magnitude are again all positive (values range from −2 to +4).

Of the 2 persons showing misfit on the TN subtest, 1 was also identified in the analysis of G2. Both have estimated abilities below the mean, but the patterns of residuals are different: for the higher-scoring of the two, the residuals are mostly negative (lowest value = −2), while for the lower-scoring person there are more positive residuals than negative ones. Thus while the former has made some unexpected errors, the latter appears to have benefitted from some chance successes.

Although this consideration of person fit in the ELTS subtest analyses has not been exhaustive, it has nevertheless drawn attention to various tendencies observed. Some of these, e.g. the intermittent large positive residuals for

low-ability persons guessing hard items correctly, might be expected to apply to multiple-choice tests in general. Others, e.g. the strings of negative residuals occurring at particular points in the item sequence, seem to relate to features of the subtests themselves.

5.3.2.3 Item Difficulty Estimates and Ability/Difficulty Scales

The Rasch difficulty estimates and standard errors for the items in each of the 8 subtests are set out in Appendix I.2. Only the final sets of difficulty estimates (i.e. those calculated after the removal of the misfitting persons discussed in the previous section) are shown in this case, since the discussion here will be concerned exclusively with these.

It will be noted that the standard errors listed for the estimated difficulties of items in G1 and G2 (see sets (i) and (ii) in Appendix I.2) are lower than those for the M1 subtests (see sets (iii)-(viii)). This, of course, results from the difference in the sample sizes; the largest standard errors are those for the smallest data set (PS).

As always, the mean item difficulty has been set to zero in each analysis. It can be seen from the summary statistics given at the foot of each set of estimates that although the standard deviations of item difficulty do not vary greatly across the subtests (SD range = 0.77 to 1.28), the ranges of difficulty spanned by the different subtests vary quite widely. Of the M1 subtests, the Life Sciences and Technology modules are shown to have the widest ranges, followed by the Medicine and Technology modules. The General Academic module has the narrowest range of difficulty of all 8 subtests.

Summaries of the distributions both of the item difficulties and of the person abilities can be found in Appendix I.5, where the ability/difficulty scales defined by these analyses (adapted from the BICAL output) are set out.

5.3.2.4 Item Fit

The item fit statistics for the ELTS subtests are listed in Appendix I.4. Although it is not possible to discuss these results in the same detail as in Chapter 4, since the test content must not be revealed, a general indication of the observed patterns of misfit can be given.

The means and standard deviations of the total fit t-statistics for the items in each subtest are shown at the foot of the tables in Appendix I.4. It can be seen

that for some subtests (e.g. ME and PS) these are close to the theoretical values of 0 and 1, while for others (e.g. G1 and GA) the standard deviations are substantially larger than this. Although one might wish to take these differences into account in deciding on limits for acceptable fit, the theoretically determined value of 2 will be used consistently here.

Numbers (and names) of items with total fit t-values of greater than 2 are shown below. The lists of names begin with the most misfitting item in each case.

G1	:	5	G111, G125, G109, G112, G132
G2	:	3	G235, G227, G233
GA	:	6	GA03, GA05, GA17, GA26, GA13, GA16
LS	:	5	LS04, LS16, LS08, LS14, LS22
ME	:	1	ME35
PS	:	1	PS06
SS	:	5	SS14, SS22, SS24, SS19, SS04
TN	:	5	TN13, TN27, TN38, TN24, TN28

For the subtests in which more than 1 item is identified as showing significant misfit, it is usually the case that 2 or more of the misfitting items occur within the same smaller item subset. The only exception to this is the LS subset, in which all the misfitting items relate to different texts. In the SS module, 4 of the items identified relate to the same text, while in both GA and TN, 2 consecutive items on the same text are found to misfit. 2 of the misfitting items in G1 belong to the same subsection (choosing the most accurate paraphrases of given sentences), and in G2 all 3 misfitting items come from the same subsection (comprehension questions on a taped 'seminar'). Thus in seeking to account for the misfit observed here, one would need in some cases to consider the idiosyncracies of particular items, and to check, for example, whether a particular distractor had exerted undue influence on higher-level testees (as seems to be the case for item G111). In other cases, one would need to examine for possible effects relating to a given item type or text.

The items listed at the beginning of each of the tables in Appendix I.4 are those showing the best (or most extreme) fit. It can be seen that some subtests, notably G1, GA and LS, contain items with large negative values for the total fit-t and large positive values for the between-group fit-t, indicating that the success rate for the lower-level persons was even lower than expected, and that for the higher-level persons even higher. In G1 and LS, the most extreme items of this kind occur either at or very near to the end of the subtest, suggesting the possible influence of a time effect. The first item listed for GA, however (GA34), seems to have shown extreme discrimination for some reason connected with its

content rather than its position in the test.

The ME, PS and TN modules contain no noteworthy cases of extreme discrimination; indeed, the lowest total fit t-values observed for these subtests were -2.06, -2.14 and -2 respectively, with corresponding between-group fit t-values as low as 1.35, 0.74 and 1.57.

Discussion of the proportions of correct answers across the 6 ability subgroups (see Appendix I.3) is not included here, since this information is presented in graphical form in Section 5.5.1.

5.3.2.5 Person Separation

The person separability index, or test reliability of person separation, is shown for each subtest at the foot of the tables in Appendix I.1. As has already been noted, this is the Rasch-based equivalent of the K-R20 reliability coefficient. The slight differences between these values and those reported in the results of the traditional analysis (see Appendix H.1) result from the changes to the data sets brought about by the removal of misfitting persons, and those with zero or perfect scores.

The person separability indices range from 0.78 (for G2) to 0.87 (for GA), and the number of person strata, i.e. distinct ability bands, into which the persons in these samples are separated by the subtests varies accordingly. In general, however, the number of person strata is approximately 3 in each case. Since these samples are not representative, it is not possible to generalise from these results to the population of ELTS candidates; it would, however, be of interest to obtain information of this kind using a representative sample, to see how well this corresponds with the number of bands currently used for the reporting of ELTS scores.

5.4 Comparison of Traditional and Rasch Analyses

The differences between traditional and Rasch analyses in terms of the kind of information yielded by each were discussed in some detail in Chapter 4. In this chapter, therefore, attention will be restricted to (a) a comparison of facility values and Rasch item difficulties, and (b) a comparison of the particular items identified by the discrimination and item fit statistics as being inconsistent with their respective sets.

5.4.1 Comparison of Facility Values and Rasch Difficulty Estimates

In order to compare the stability of the facility values and Rasch difficulty estimates for some of the ELTS items, 2 data subsets containing (i) the 500 highest-scoring persons and (ii) the 500 lowest-scoring persons were drawn from the complete data sets for G1 and G2. The means, standard deviations and ranges of the raw scores for the subgroups were as follows:

	Mean	SD	Range
G1 High scorers	35.8	2.2	32-40
G1 Low scorers	20.3	3.9	7-25
G2 High scorers	29.2	2.1	25-34
G2 Low scorers	17.4	3.2	4-21

Facility values and Rasch difficulty estimates were obtained separately for each of the groups, and the pairs of values plotted as in Chapter 4. The results for G1 are shown in Figures 5.1 and 5.2, and those for G2 in Figures 5.3 and 5.4; the lists of values can be found in Appendix H.4 (for the facility values) and in Appendices J.8 and J.9 (for the Rasch difficulty estimates).

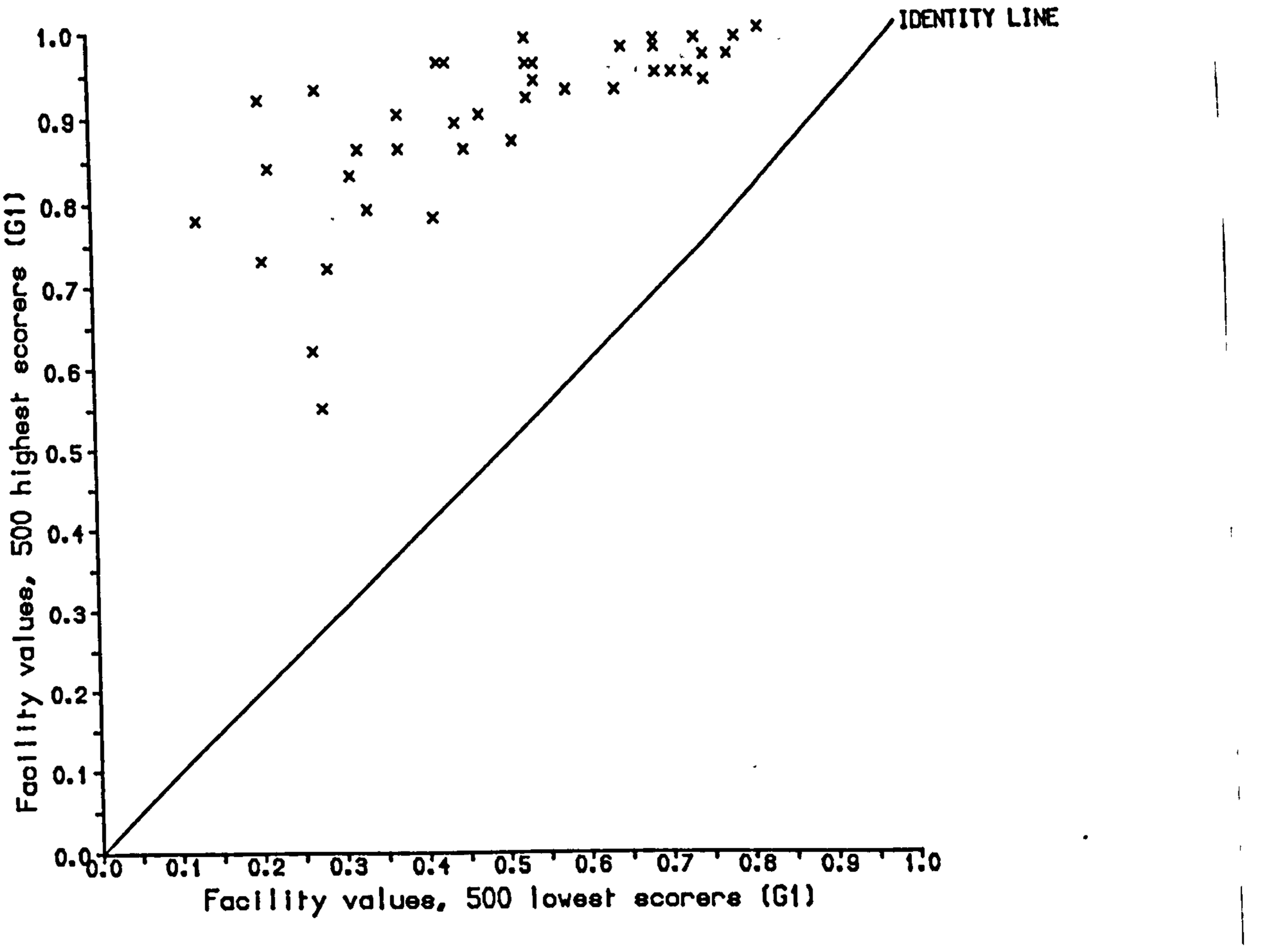


Figure 5.1 G1 Facility Values, High vs Low Scorers

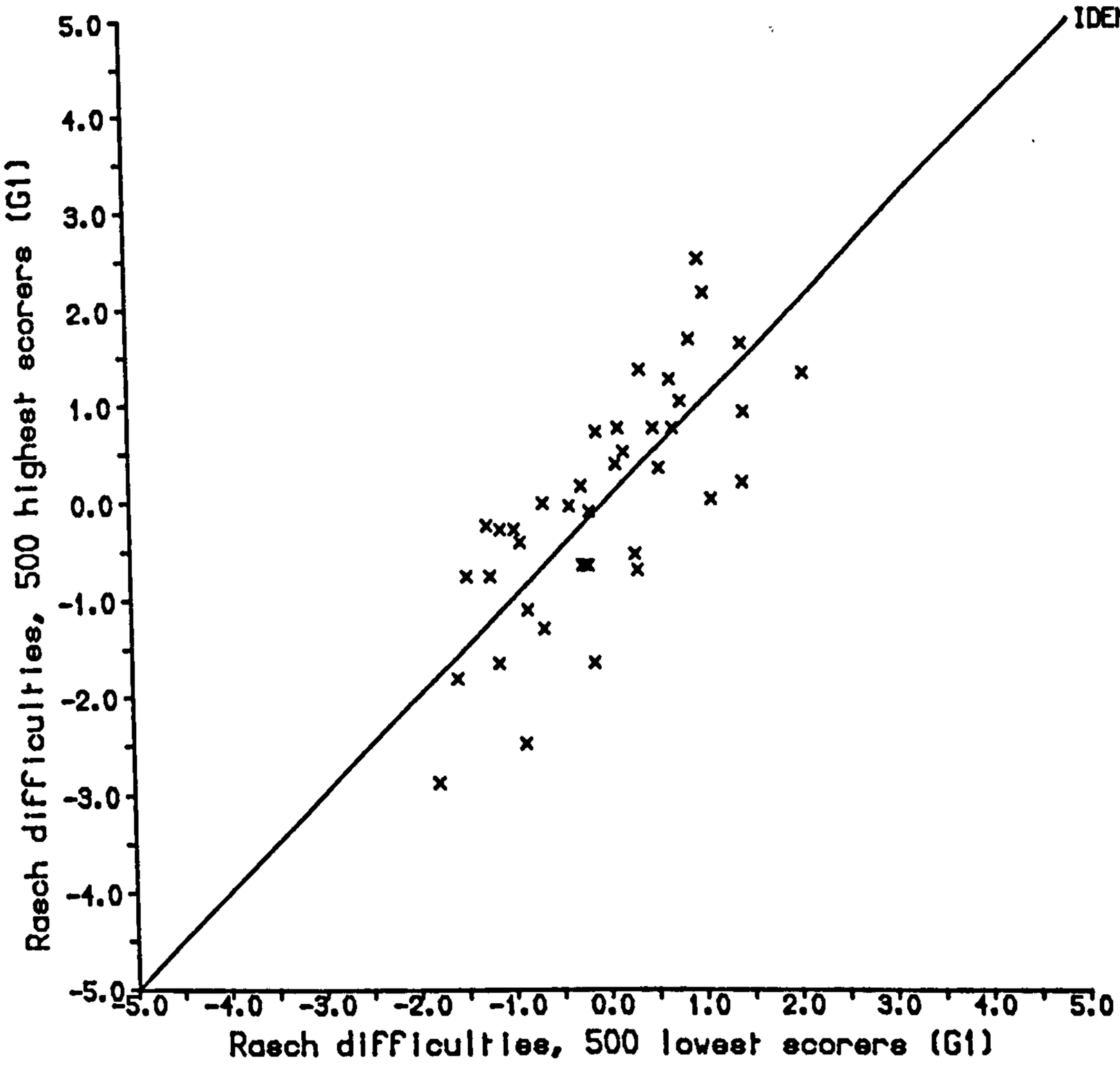


Figure 5.2 G1 Rasch Difficulty Estimates, High vs Low Scorers

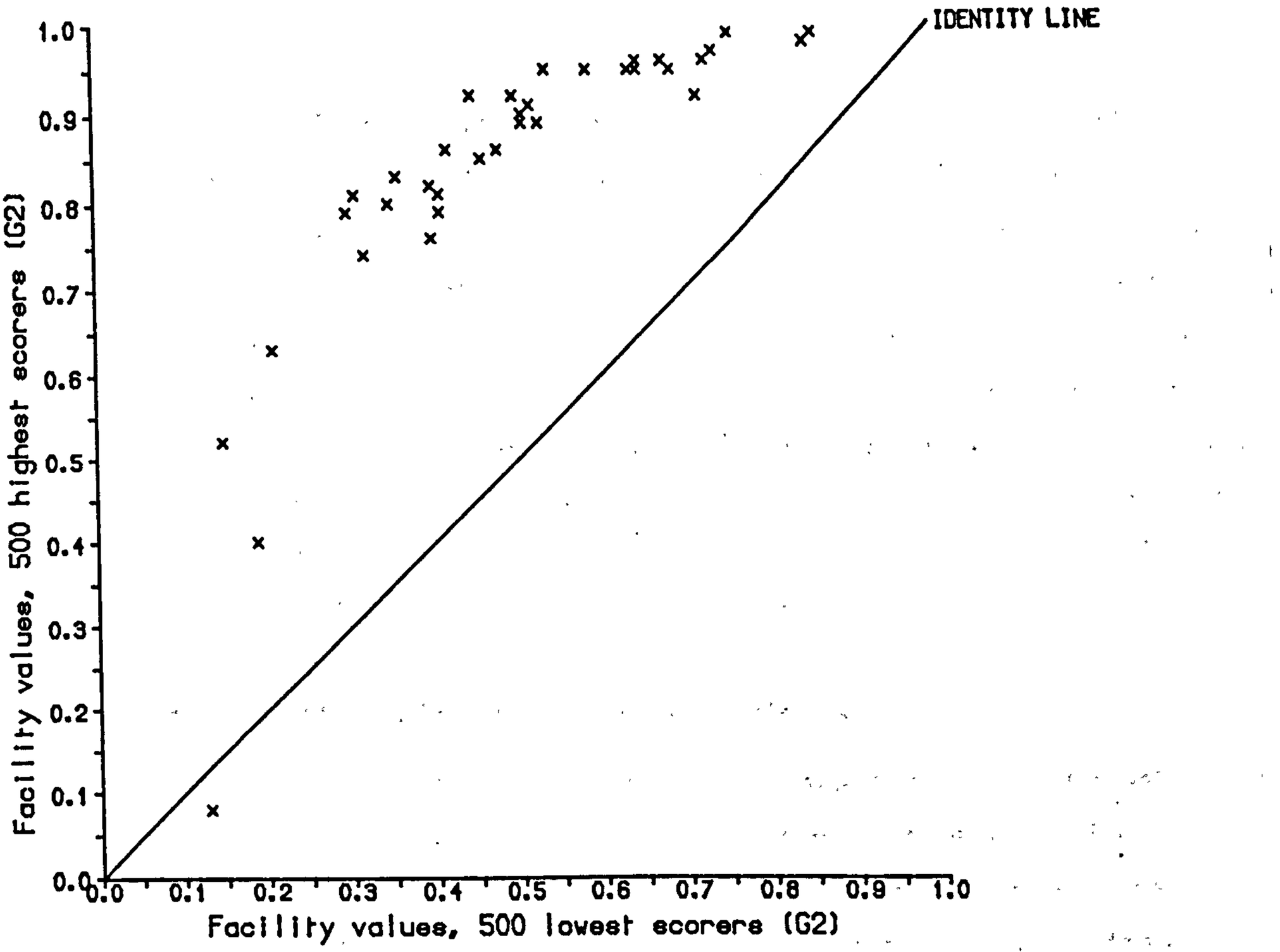


Figure 5.3 G2 Facility Values, High vs Low Scorers

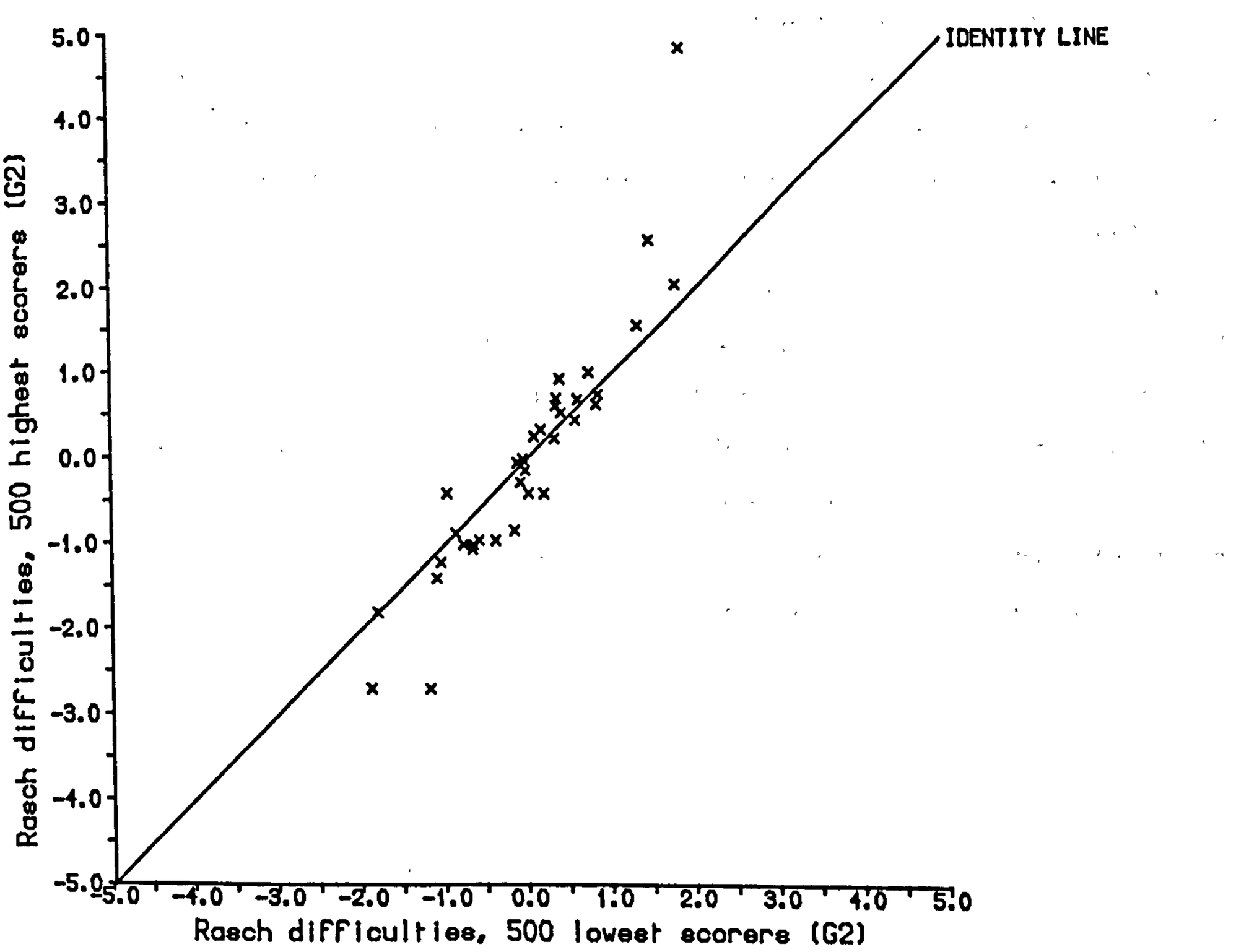


Figure 5.4 Rasch Difficulty Estimates, High vs Low Scorers

It can be seen from Figures 5.1 to 5.4 that, as was also demonstrated in Chapter 4, the Rasch difficulty estimates show considerably greater stability between the two subgroups in both cases. Even though the subgroups used in these comparisons differed less widely in score levels than those used in Chapter 4, the plotted facility values for both G1 and G2 nevertheless cluster in the upper part of the two graphs. The corresponding Rasch difficulty estimates, on the other hand, are, in general, spread along the identity line; indeed, in the case of G2, most of the pairs of estimates appear to correspond very closely for the two groups. The question of whether the two sets of estimates obtained for each subtest are statistically equivalent, however, will be dealt with in Section 5.5.3.

5.4.2 Comparison of Discrimination and Fit Statistics

Table 5.2 below shows the items identified as misfitting (total fit t-value > 2) in the various ELTS subtests, together with their point biserial coefficients and their rank position by point biserial. The list of items begins in each case with the most misfitting item; a rank position of 1 for the point biserial denotes the lowest value observed for that subtest.

Table 5.2 indicates that the subtests vary in the degree of consistency shown between the rankings by total fit t-values and point biserial. For G2, SS and TN, for example, the most questionable items identified by each index correspond very closely, even though they may not be ordered in quite the same way. For ME and PS, too, the misfitting items correspond with low rank by point biserial (second lowest in each case); the items with the lowest point biserials did not show significant misfit, but were ranked either 2nd or 3rd by total fit-t. In the case of G1, GA and LS, however, a slightly different pattern is observed: although there is some degree of correspondence in the two rankings, at least two of the misfitting items in each case were not among those ranked lowest by point biserial. The subtest for which this is most noticeable is G1, where for example the 5th least well fitting item was only the 17th lowest by point biserial.

Subtest Items	Most Misfitting Items	Point Biserial	Point Biserial Rank
G1	G111	0.17	1
	G125	0.25	5=
	G109	0.25	5=
	G112	0.30	11=
	G132	0.34	17=
G2	G235	0.13	2
	G227	-0.15	1
	G233	0.22	4=
GA	GA03	-0.08	1
	GA05	0.12	2
	GA17	0.20	4
	GA26	0.26	6=
	GA13	0.29	9=
	GA16	0.30	11
LS	LS04	0.02	2
	LS16	0.10	4
	LS08	0.05	3
	LS14	0.17	7=
	LS22	0.19	12=
ME	ME35	0.15	2=
PS	PS06	0.27	2
SS	SS14	0.10	2=
	SS22	0.12	4=
	SS24	0.16	6
	SS19	-0.04	1
	SS04	0.10	2=
TN	TN13	-0.10	1
	TN27	0.07	2
	TN38	0.22	7
	TN24	0.17	4=
	TN28	0.14	3

Table 5.2 Most Misfitting and Least Discriminating ELTS Items

If one considers the facility values of some of the items ranked low by point biserial but not appearing in Table 5.2, the relationship between point biserial and item difficulty again becomes apparent. For example, the item with the lowest point biserial in LS (LS35) has a facility value of only 0.05, and thus will have shown low discrimination as a result of extreme difficulty. The total fit t-value for this item, however, is relatively low (0.18), and corresponds to a rank of 17th

by degree of misfit. The item ranked 4th by point biserial in G1 (G119), on the other hand, appears to have shown low discrimination because of its extreme easiness for the group (facility value = .9); the total fit t-value in this case is -0.58, corresponding to a rank position of 18th. Thus there are again instances where the particular items identified by the traditional discrimination and the Rasch fit statistics differ as a result of the fact that the former are related to item difficulty while the latter are not.

In general, however, the effect of this is less marked in the results of the ELTS analyses than in those of the cloze-type test analyses reported in Chapter 4. This is explained by the difference between the data sets in terms of the ranges of item difficulty and person ability spanned: in the case of the ELTS data, both the item sets and the person samples are more homogeneous in level, with the result that fewer items appear at the extremes of easiness or difficulty.

5.5 Rasch Analysis of ELTS Data: Further Investigations

In this section, checks of the type described in Section 4.5 are applied to the ELTS data, to provide a point of comparison for the results reported there, and to investigate further the nature of the ELTS data.

5.5.1 Observed vs Expected ICCs

The estimated observed ICCs (i.e. the proportions of correct responses across the 6 ability subgroups, listed in Appendix I.3) are plotted for the items in the ELTS subtests in Figures 5.5 to 5.11. Within each subtest, the subsets of items either of a particular type, or relating to a particular text, have been plotted in separate groups, so that any patterns to be found either within or across such subsets are more easily observable. For the M1 subtests, the items relating to bibliography and index material (and, in one case, glossary material) have been treated as a single subset in each case. For G1 and G2, details of the types of item contained in the various subsets may be found by referring back to Section 5.1.1.

Since these subtests together involve a large number of items, the individual results are not discussed in detail; the purpose here is rather to provide a visual summary of the observed response patterns across ability groups. Particular items can, however, be identified by their sequence of y coordinates, which appear in the 'Item Characteristic Curve' sections of Appendix I.3.

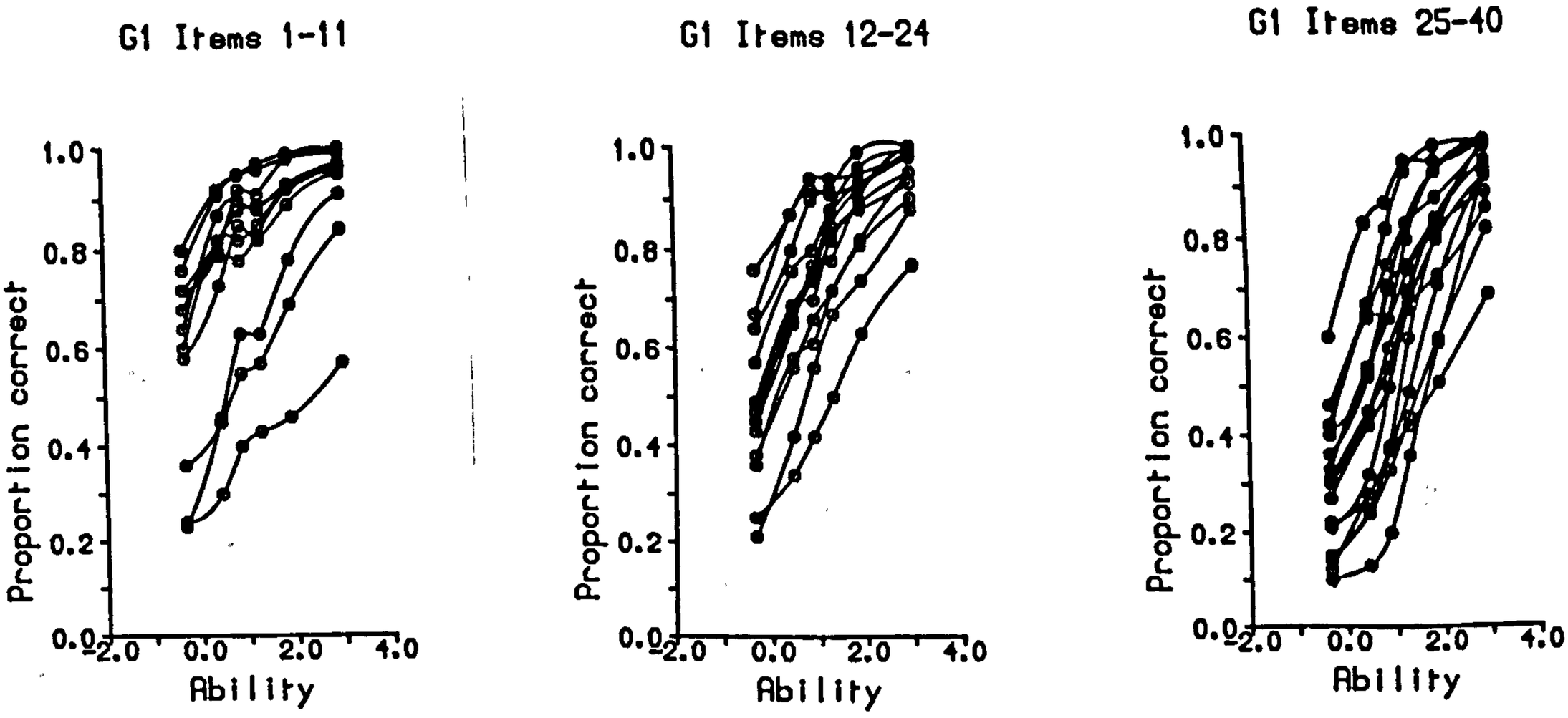


Figure 5.5 Observed ICCs for G1

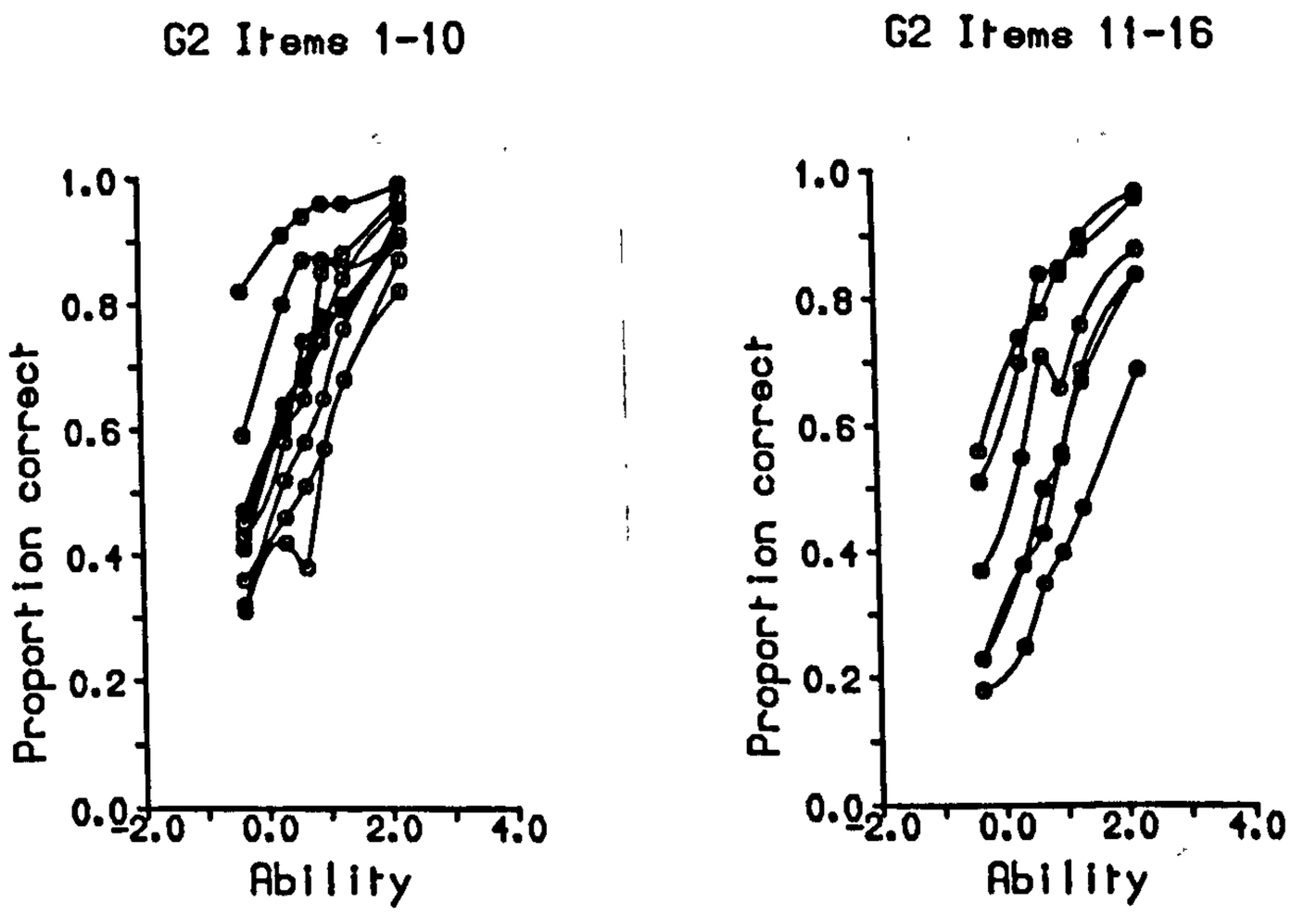


Figure 5.6 Observed ICCs for G2

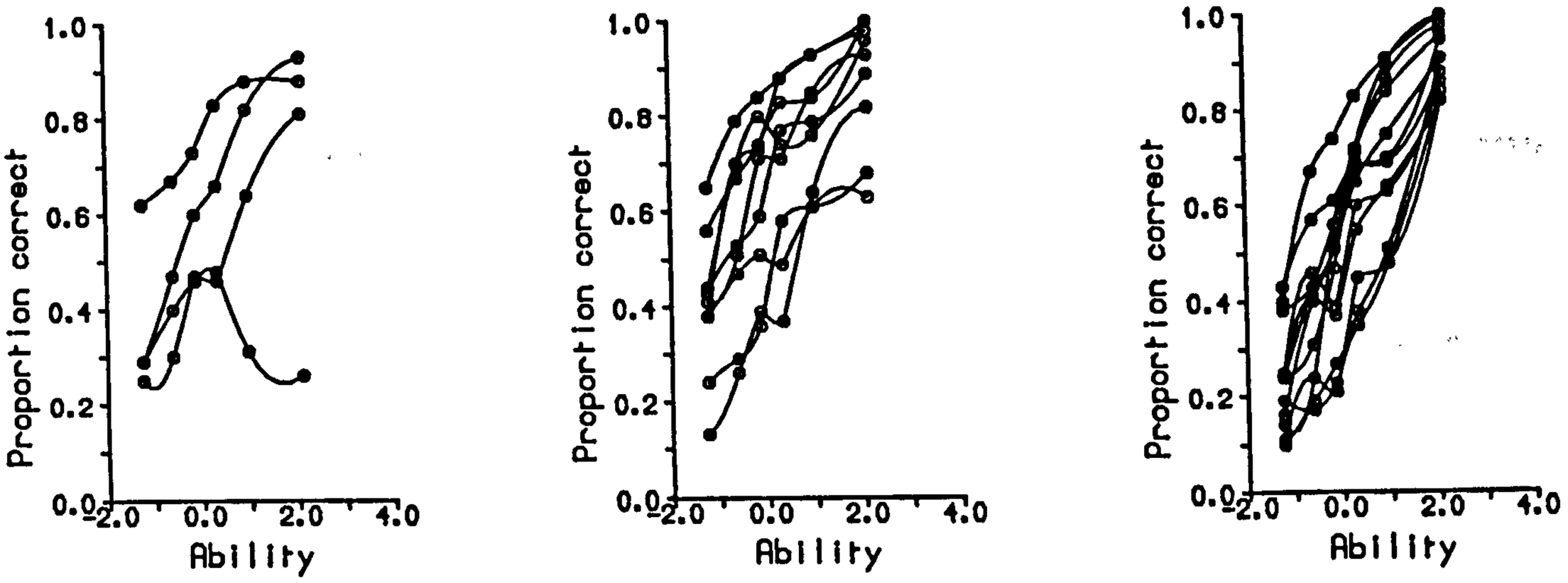


Figure 5.7 Observed ICCs for GA

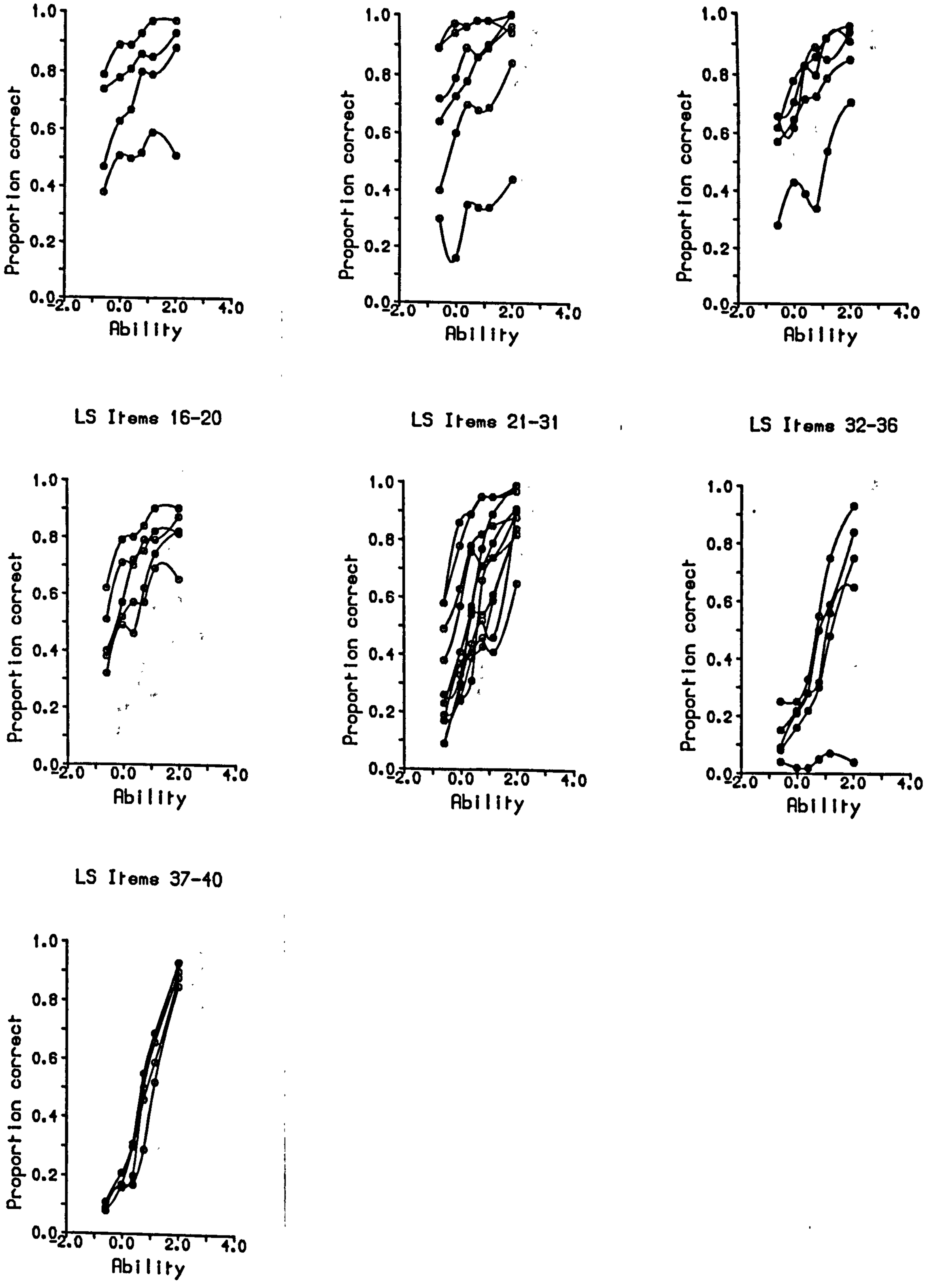


Figure 5.8 Observed ICCs for LS

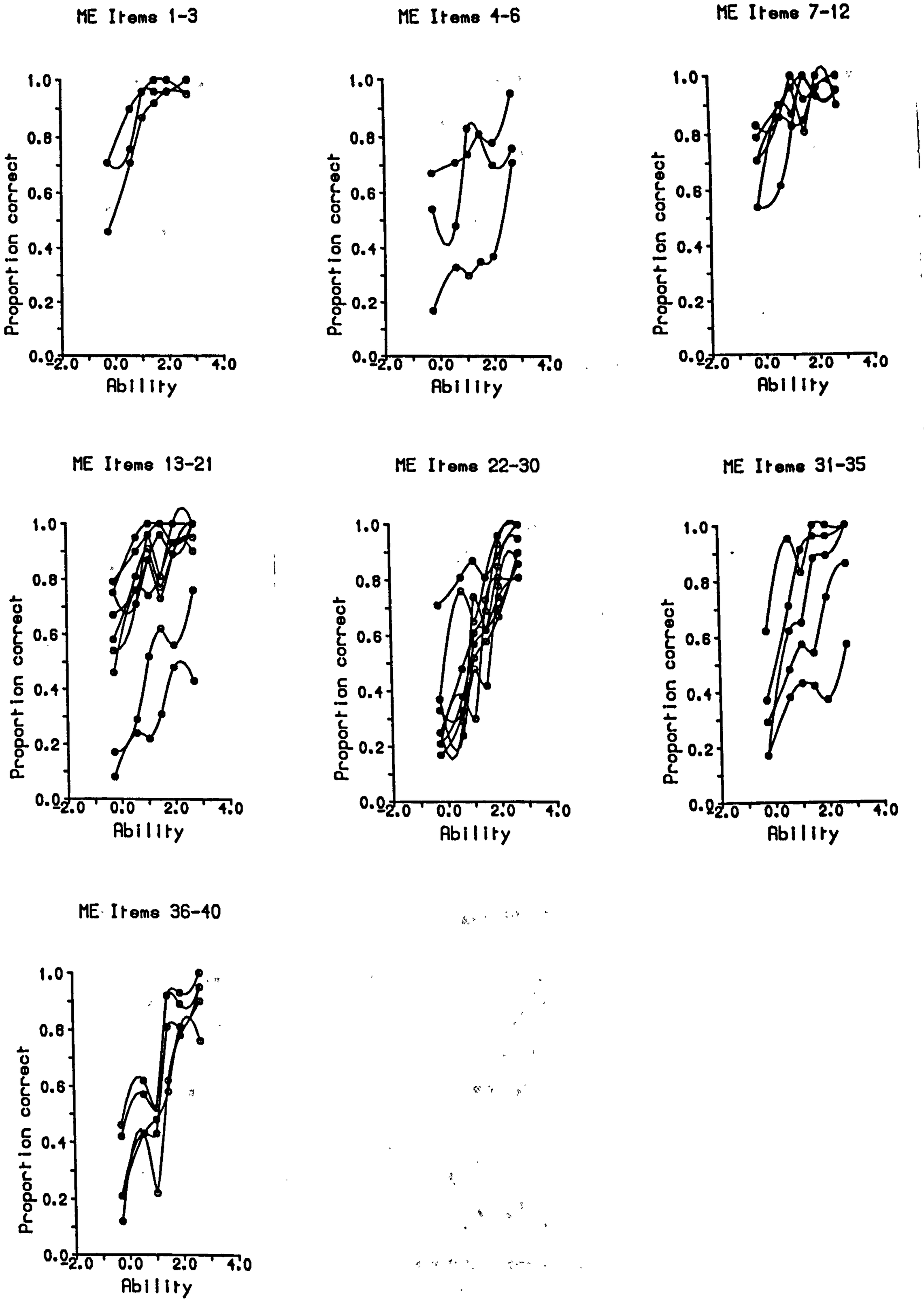


Figure 5.9 Observed ICCs for ME

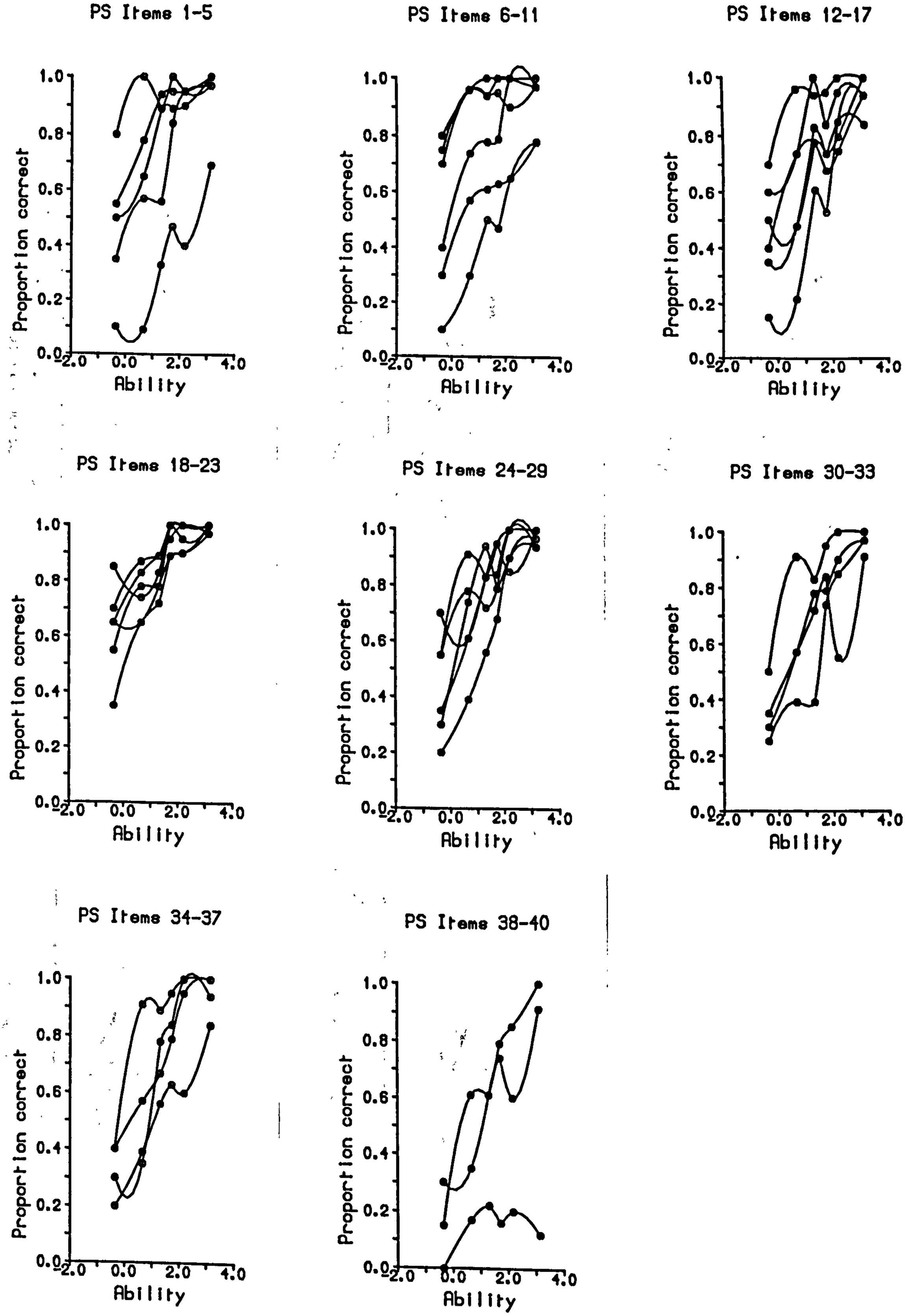


Figure 5.10 Observed ICCs for PS

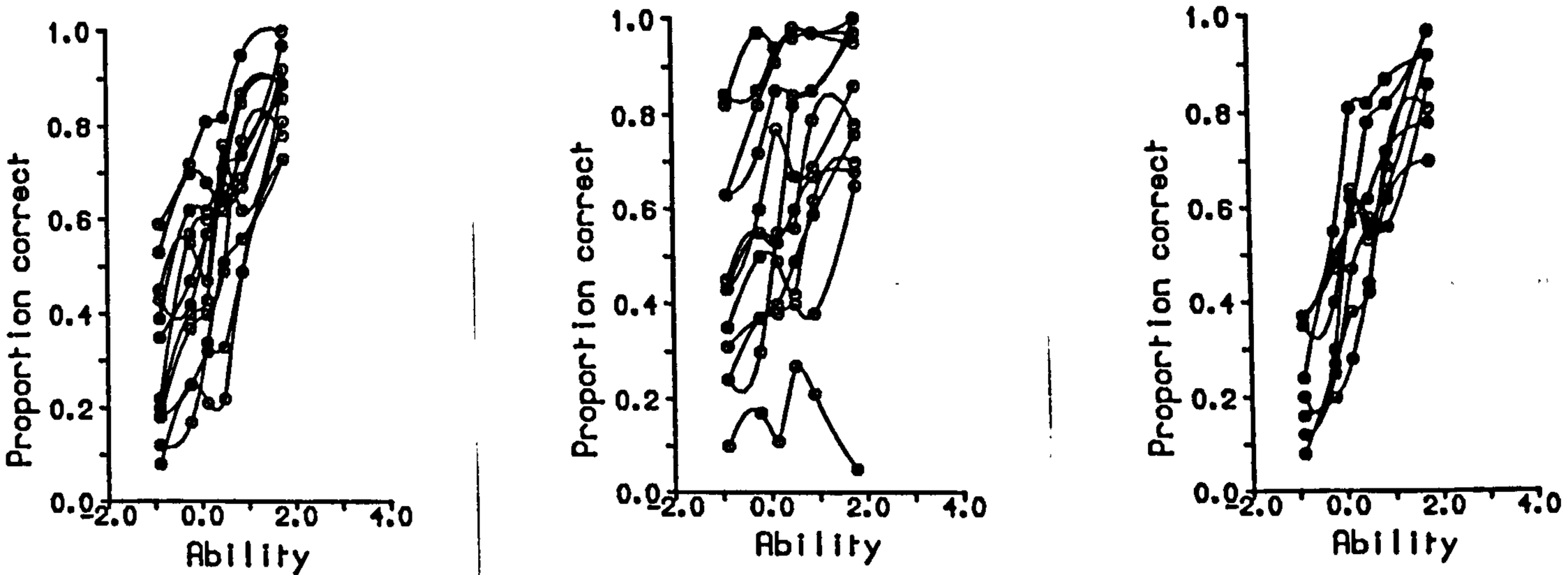


Figure 5.11 Observed ICCs for SS

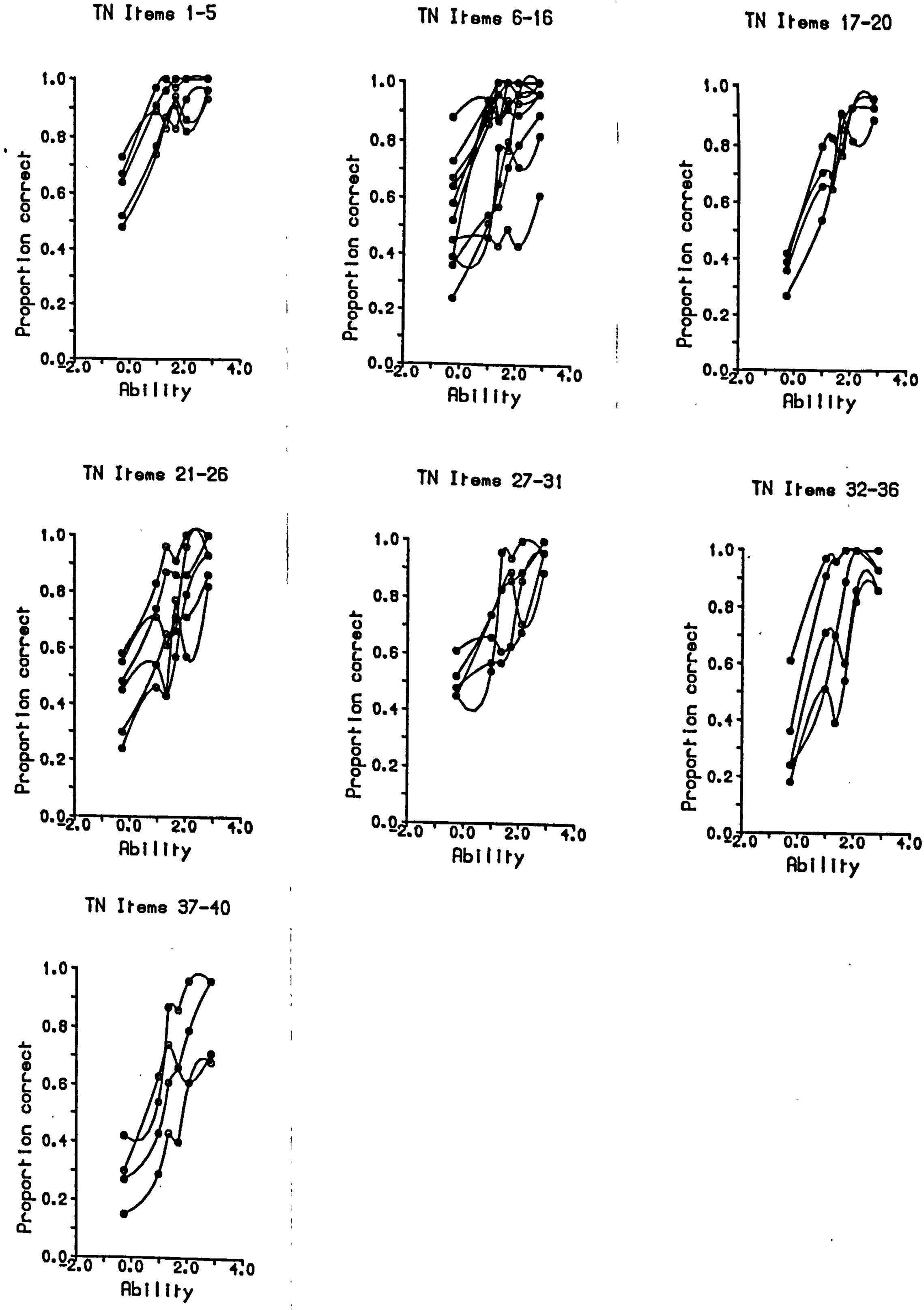


Figure 5.12 Observed ICCs for TN

It should be noted that the score groups into which the testees have been divided in these analyses in many cases span very narrow ranges (see foot of tables in Appendix I.3). Indeed, in view of the measurement error which will have entered into the scores, it is likely that some of the subgroups at best represent only rough groupings by 'true' ability. This point must be borne in mind in considering the shape of these estimated ICCs, particularly those for some of the Modular subtests, where numbers in each subgroup are also relatively small.

5.5.1.1 Conformity between Data and Model

The estimated ICCs for G1 (see Figure 5.5) in general show the expected increase across the 6 ability subgroups, though they vary in their steepness. The steepest curves are for items in the 3rd item subset in G1; the gradual increase in difficulty and discrimination observed across the 3 item subsets may in part be due to time effects. In order to assess the extent to which this apparent progression reflects increasing item difficulty, one would need to vary the order of administration, or to ensure that all testees had sufficient time to attempt all questions.

For G2 (see Figure 5.6) there appears at first sight to be less consistency in the shape of the curves than for G1. However, on closer inspection it can be seen that this effect is created largely by a small number of items with decreasing proportions correct in the area of the 2nd-4th subgroups, where subgroup score ranges are in any case very narrow. A particularly odd pattern, however, occurs in the last of the 4 item subsets in G2, where one item was answered correctly by a higher proportion of those in the 1st ability subgroup (raw score range 1-18) than in the 6th subgroup (raw score range 28-34). This curve is that of item G227, which was identified both by the traditional and Rasch indices of item quality as being the most inconsistent in the subtest.

The most noticeable departure from expectation in the plotted curves for GA appears in the 1st graph in Figure 5.7, where the ICC for one item is almost bell-shaped. This again corresponds to the most questionable item identified both by the fit statistics and the discrimination indices (item GA03). Looking at the pattern of curves across all 6 item subsets in GA, one notes that the curves become particularly steep at the end of the test; this may again be attributable, at least in part, to the influence of the time limit. The same effect is also noticeable in the curves for the LS items (see Figure 5.8).

The person samples for ME, PS and TN were the smallest in the ELTS data set

used here (142, 132 and 182 respectively), and thus subgroup sizes for these subtests were relatively small, ranging from 19 to 35. Small differences in the numbers answering correctly will clearly have a greater effect on the proportions than when subgroup sizes are larger, with the result that the subgroup differences reflected in the plotted curves appear somewhat exaggerated (see Figures 5.9, 5.10 and 5.12). As was mentioned in Chapter 4, the plotting method used can also be seen in some cases to have contributed to the irregularity in these curves: a particularly noticeable example appears in the lower part of the 5th graph in Figure 5.12 (TN, items 27–31).

The greatest irregularity in the ICCs for the SS items can be seen in the 2nd graph in Figure 5.11; this item subset contains 4 of the 5 items identified as showing significant misfit in this subtest. The curve in the lower part of the graph is that of item SS19, which, although not the most misfitting item in terms of total fit–t, showed the greatest misfit between groups. The items positioned towards the end of the subtest appear more difficult in general than those occurring earlier. This again suggests the operation of a time effect, though in this case one which may have affected the higher–level candidates as well as the lower–level ones.

Direct comparison of the curves shown for the different Modular subtests may not be appropriate in view of the differences in samples and sample sizes. However, within the subtests it is possible to discern certain patterns (e.g. relating to possible time effects), and to note whether irregularities occur throughout the subtest or within a particular cluster of items.

5.5.1.2 Variation in Discrimination

The Rasch–based discrimination indices for the items in each subtest are listed with the fit statistics, in Appendix I.4. As was explained in Chapter 4, this index reflects the difference in steepness between the observed and expected ICC for each item; the closer the value to 1, the greater the correspondence between the two.

Numbers of items with Rasch–based discrimination indices in each of 5 intervals are shown for each subtest in Table 5.3 below. (G2, it will be remembered, contains only 35 items, while all other subtests contain 40.)

Rasch-based Discrim. Index	NO. OF ITEMS							
	G1	G2	GA	LS	ME	PS	SS	TN
.39 or less	1	1	2	3	1	1	1	2
.4 to .69	2	1	4	7	3	1	6	3
.7 to .99	15	10	11	12	16	17	11	10
1 to 1.29	16	21	16	7	12	17	14	16
1.3 or over	6	2	7	11	8	4	8	9

Table 5.3 Summary of Rasch-Based Discrimination Indices for ELTS Items

It is clear from Table 5.3 that for the majority of items in each subtest, there is quite close correspondence between the observed and expected curves. However, each subtest also contains a number of items which, at least for the person samples used here, have either flatter or steeper ICCs than expected. The subtests which are shown to depart least from expectation are G2 and PS, for which numbers of items with indices departing substantially from 1 in either direction are small. From the figures presented here, LS appears to deviate the most from expectation, with roughly half its items showing either lower or higher discrimination than expected.

Thus while there is in general a fairly close correspondence between observed and expected discrimination, the 8 data sets under consideration here show varying degrees of departure from the assumption of equal discrimination.

5.5.2 Dimensionality of the Data

5.5.2.1 Guessing and Time Effects

It is, of course, impossible to determine the extent to which these data sets have been affected by chance success, though at least some evidence that this had occurred was noted in some of the patterns of standardized residuals mentioned in Section 5.3.2.2. This is as one would expect in multiple-choice tests such as these. However, not all persons who could not answer (or failed to reach) certain items chose answers at random, as can be seen from the patterns of omissions for each subtest. These patterns are summarised below, since they also provide evidence of the extent to which the data may have been affected by the time limits imposed.

G2 differs from the 7 other subtests examined here in that the pace of answering is dictated by the tape on which the listening material is recorded

rather than by the reading speed of the candidate. Inspection of the numbers of persons omitting each item indicates that G2 is the only subtest in which the largest number of omissions is not either at, or very near to, the end of the test. The most frequently omitted item was in this case item G230, which was omitted by roughly 16% of the sample.

For G1 and all the Modular subtests the number of omissions is largest at the end. There are, however, striking differences in the proportions of the samples omitting items, and in the point at which numbers begin to increase.

The patterns for G1, GA and LS are similar: numbers omitting are relatively small for the first half of the items, but show a steady increase after that point, with a sharp rise for the last 8-10 items. For both G1 and GA the item most frequently omitted is the last in the subtest, while for LS it is the third item from the end. The proportions of persons omitting these items are similar for GA and LS (approximately 40% and 38% respectively), but lower for G1 (approximately 20%). SS shows a somewhat similar pattern, though the increase in omissions begins at a later point. In this case the last item was omitted by just over 30% of the sample.

For ME, PS and TN, numbers omitting are very small until the last 5 or 6 items in each case. The proportions involved again vary, however: in ME, the last few items were omitted by up to 21% of the group, while for both PS and TN the proportions omitting never exceeded about 12%; indeed, in the case of TN, only the last 6 items were omitted by more than 5% of the sample.

Of course, the omission of items may occur for reasons of item difficulty rather than lack of time, and candidates who run out of time may make rapid guesses rather than leave answers blank. However, the patterns of omissions observed here would tend to suggest that the steepness of the last few ICCs in some subtests (noted in Section 5.5.1) might indeed be attributable to time effects. It would also appear that the 8 data sets vary in the extent to which speed may have influenced the measures obtained, and thus in the degree to which they depart from unidimensionality in this respect.

5.5.2.2 Subtests Treated Singly and in Combination: Comparison of Difficulty Estimates

The ELTS subtests have so far been treated as separate tests. The investigations reported in this and the following two sections, however, are based on comparisons of the results obtained from these separate analyses with those

obtained from analyses of the subtests combined in various ways.

Since the ELTS test contains separate components for 'general' and study-related proficiency, and, within the 'general' component, for reading and listening, these divisions were treated as forming potentially separate dimensions for the purposes of the checks carried out here. Bejar's (1980) method for investigating the dimensionality of data was thus applied by (a) comparing the difficulty estimates obtained for the two General subtests calibrated separately with those obtained for the same two subtests calibrated together, and (b) comparing the difficulty estimates obtained for each of the Modular subtests calibrated separately with those obtained for the same items when calibrated with the two General subtests. The difficulty estimates used in each case were those arrived at after the removal of any misfitting persons.

The difficulty estimates from the combined analysis of G1 and G2 are listed in Appendix J.1, and those from the combined analyses of the Modular and General subtests in Appendices J.2 to J.7. The difficulty estimates from the separate subtest calibrations are plotted against these new values in Figures 5.13 to 5.20.

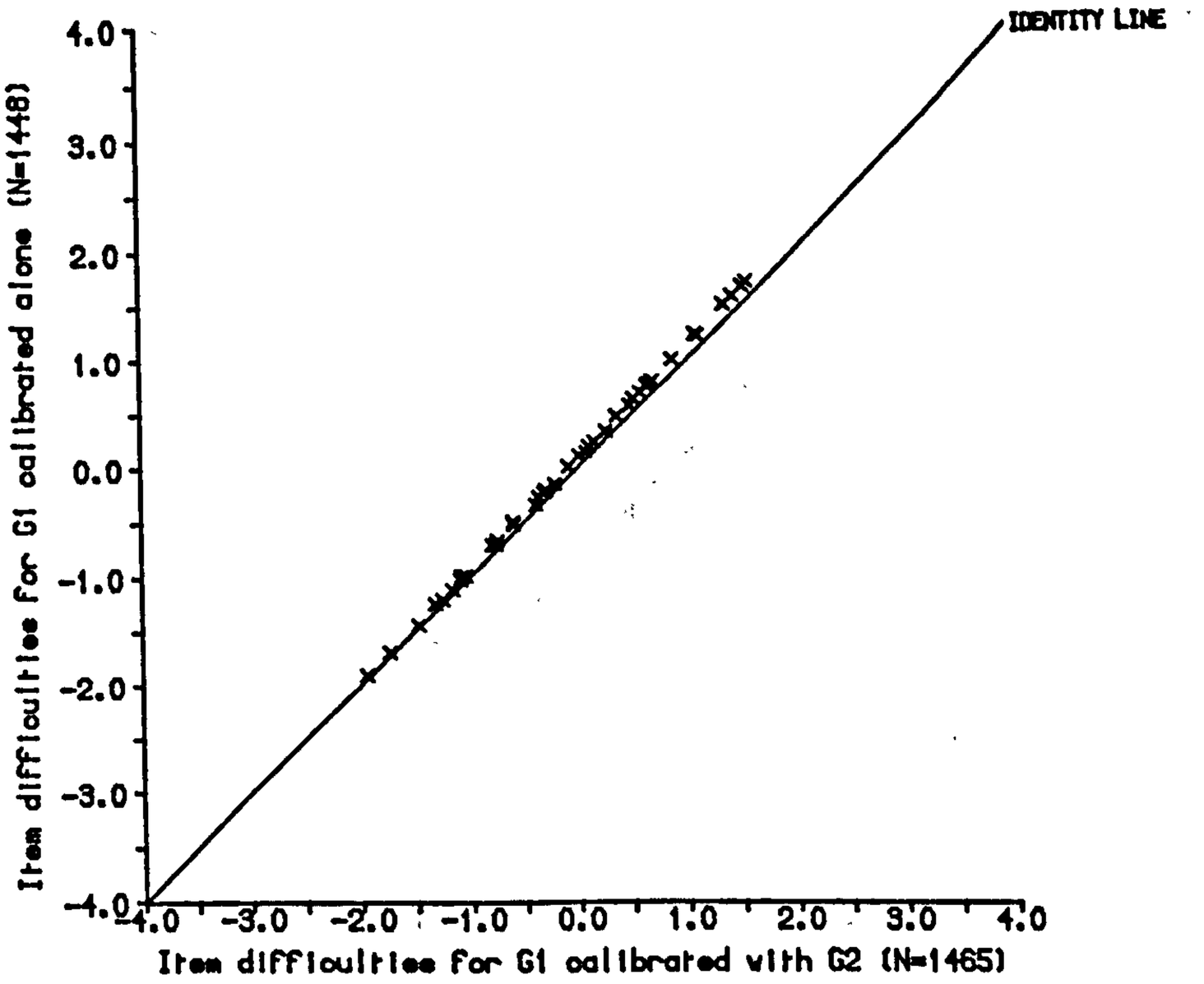


Figure 5.13 Rasch Difficulties for G1, Separate vs Combined Calibration

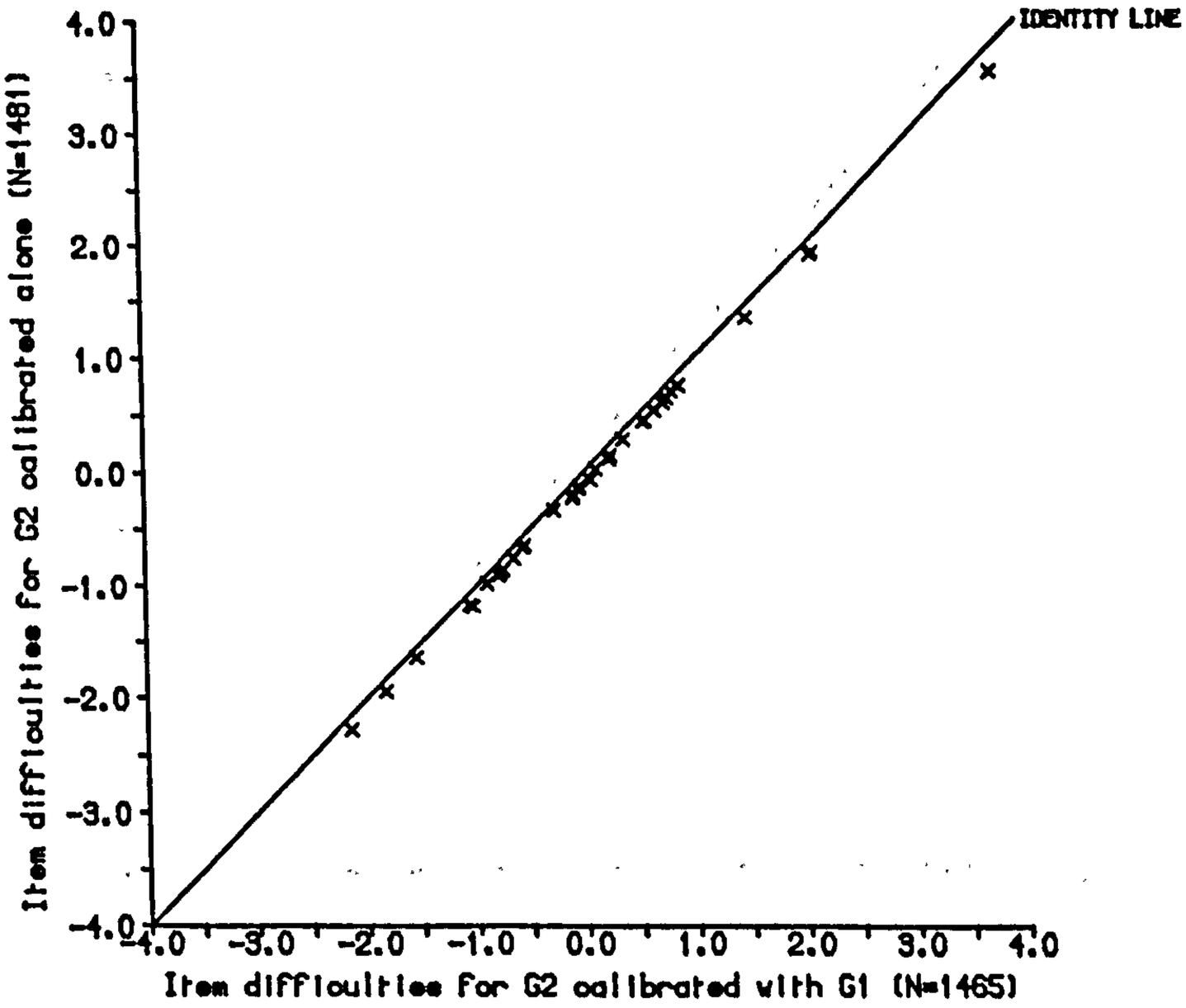


Figure 5.14 Rasch Difficulties for G2, Separate vs Combined Calibration

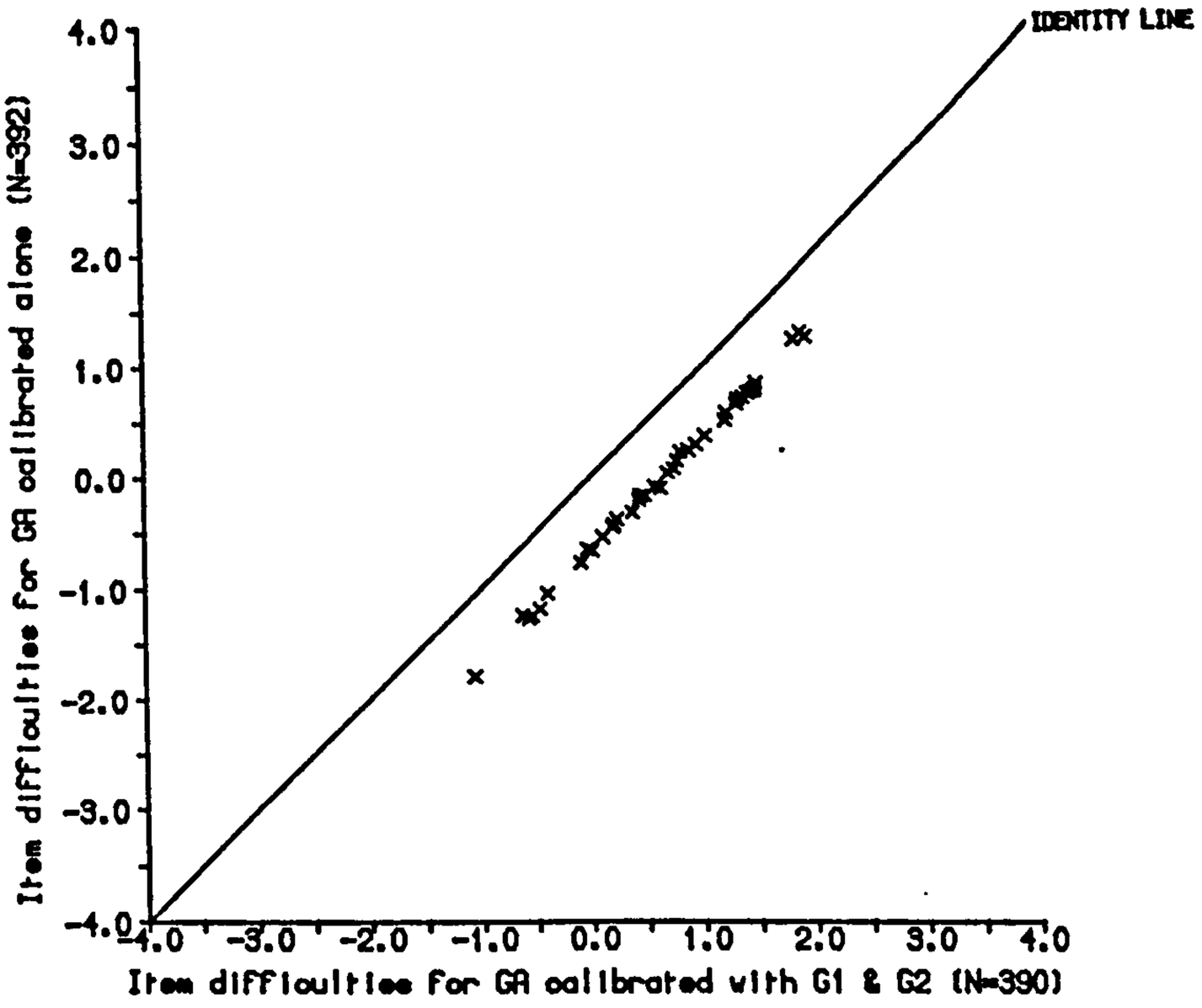


Figure 5.15 Rasch Difficulties for GA, Separate vs Combined Calibration

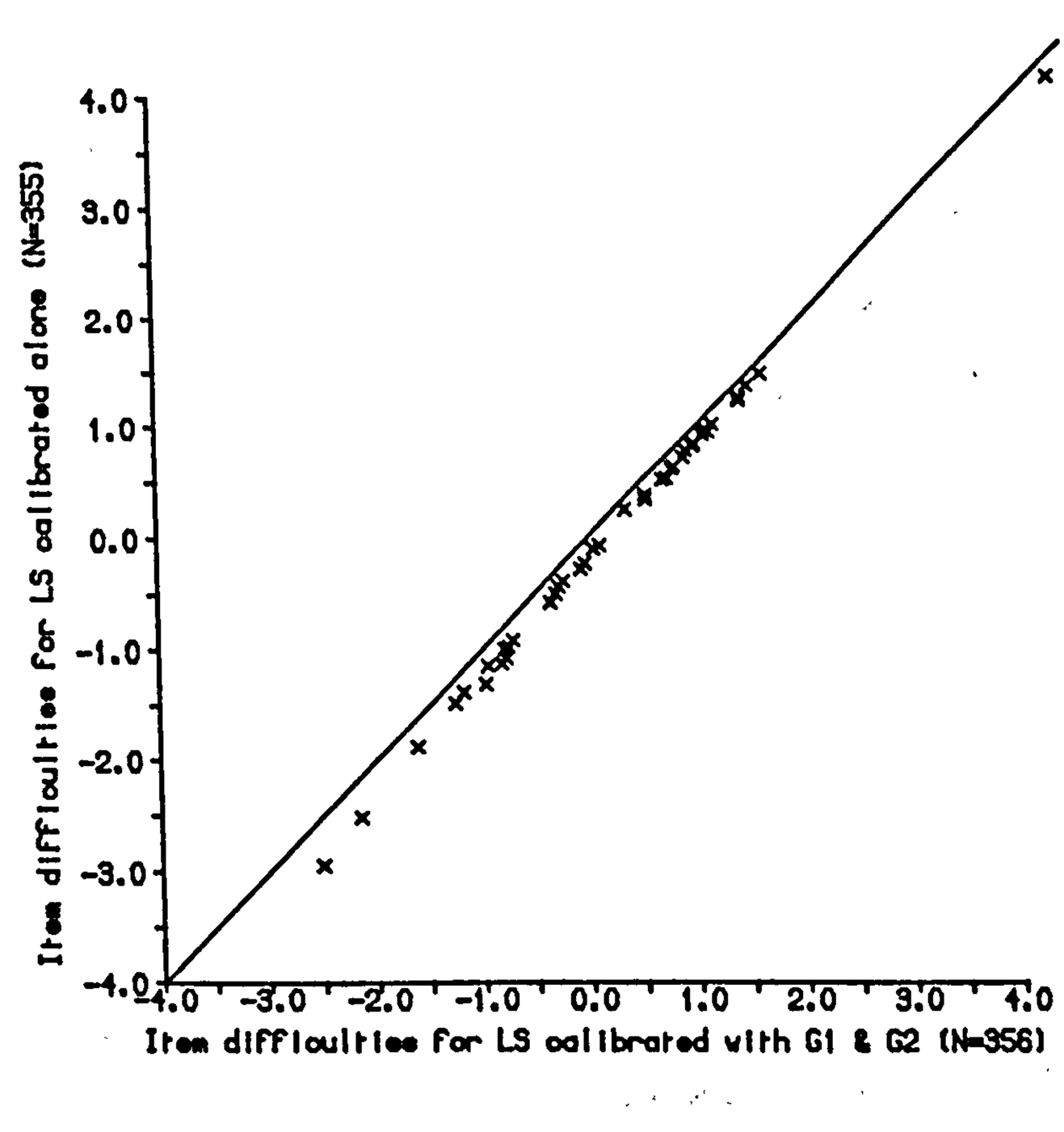


Figure 5.16 Rasch Difficulties for LS, Separate vs Combined Calibration

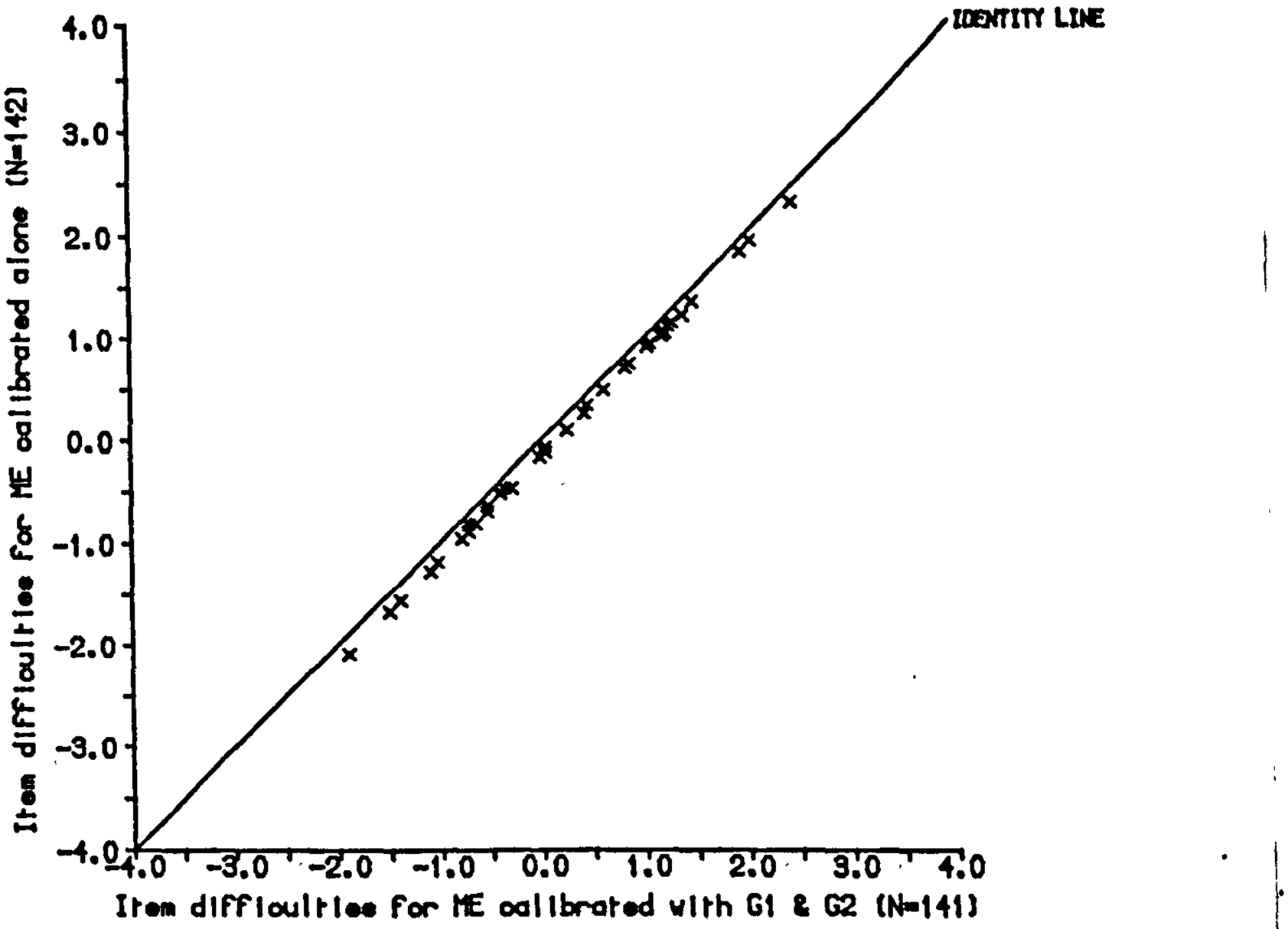


Figure 5.17 Rasch Difficulties for ME, Separate vs Combined Calibration

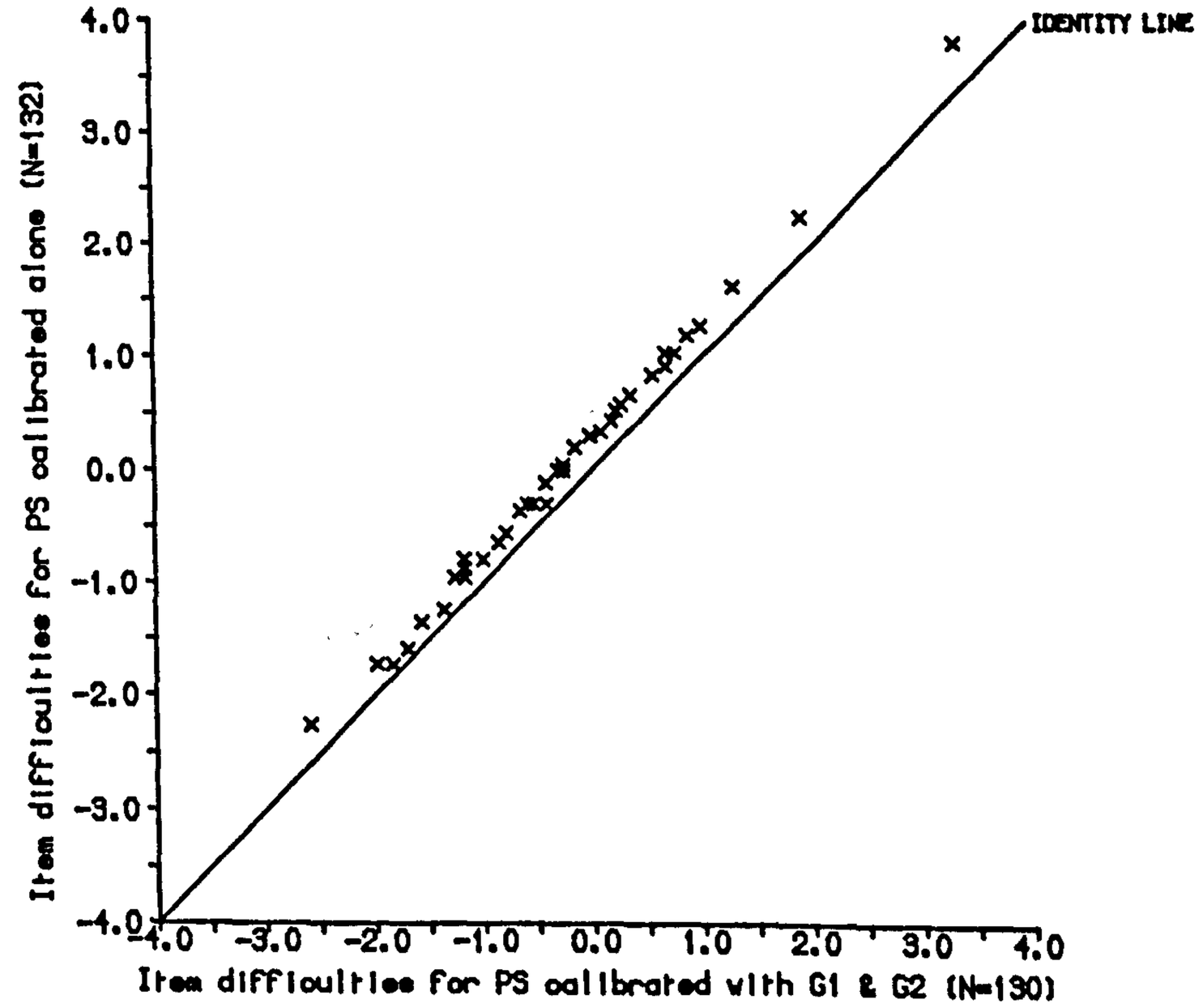


Figure 5.18 Rasch Difficulties for PS, Separate vs Combined Calibration

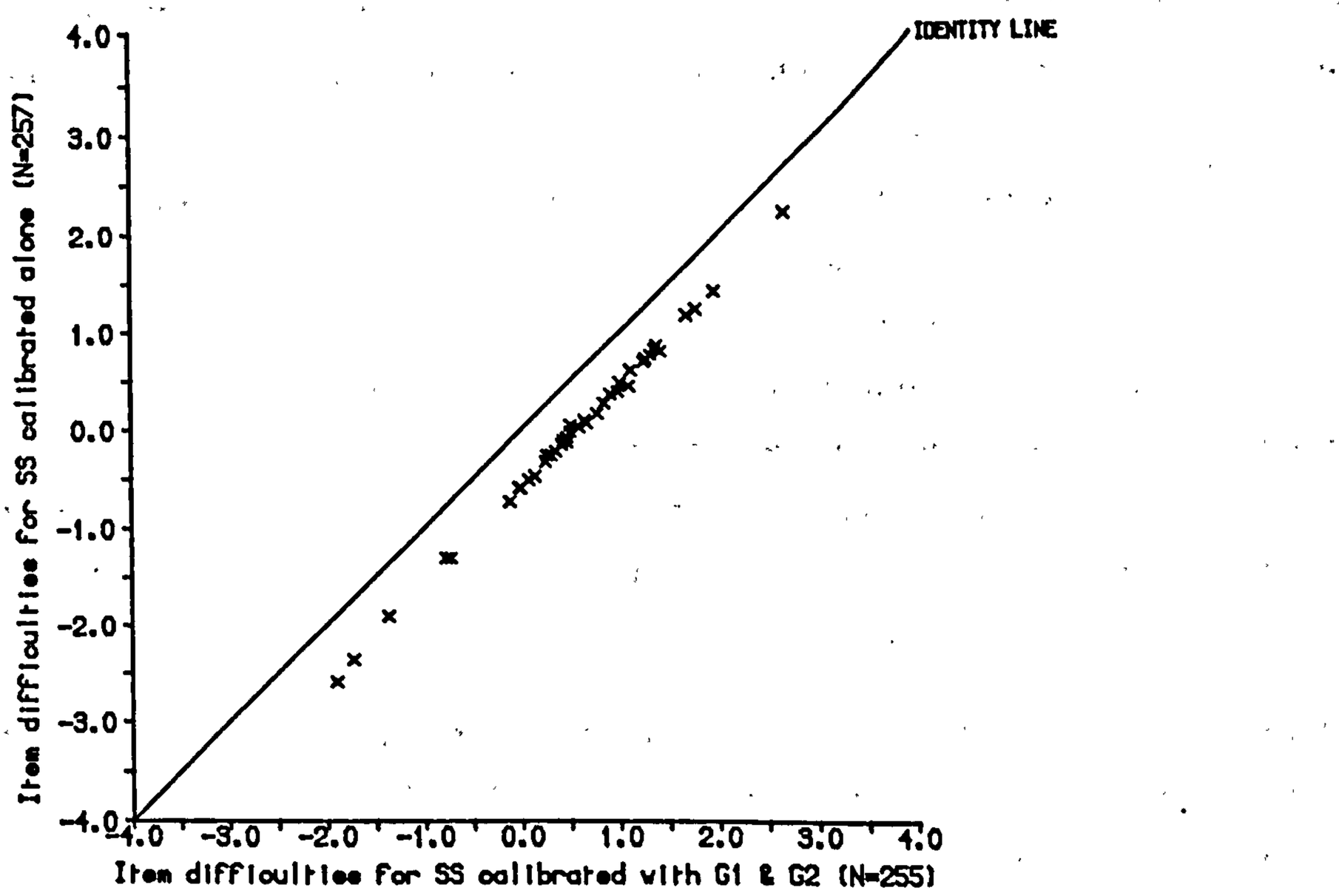


Figure 5.19 Rasch Difficulties for SS, Separate vs Combined Calibration

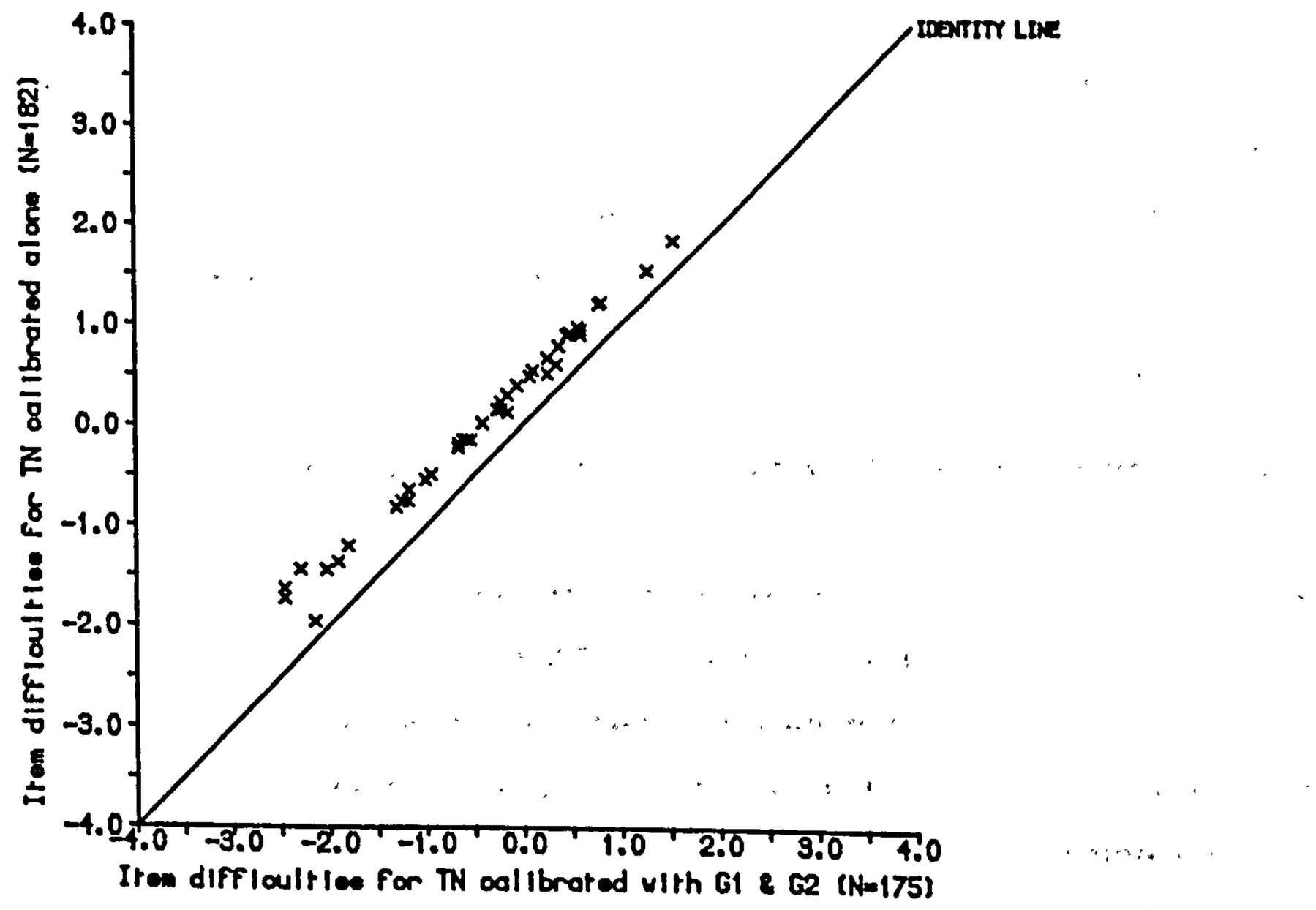


Figure 5.20 Rasch Difficulties for TN, Separate vs Combined Calibration

It can be seen from Figures 5.13 and 5.14 that the difficulty estimates obtained from the combined calibration of the two General subtests differ very little from those obtained from the separate calibrations: in both graphs, the points form a line very close to the identity line, and almost parallel with it. No adjustments were made to the sets of estimates to centre them around the same mean for each comparison, and thus the closeness of all the points to the identity line in both figures reflects the similarity of the G1 and G2 subtests in terms of their difficulty. According to Bejar's (1980) interpretation, these results would also indicate that the two subtests measure on a single dimension; judgement on this point is reserved until Section 5.5.2.3, however, when comparisons of person abilities from the same separate vs combined subtests are carried out.

The lines of points for the Modular vs combined General and Modular calibrations (see Figures 5.15 to 5.20) are again almost parallel with the identity line in each case. They vary in their distances from the identity line, however, and in terms of whether they fall above or below it. Since no adjustments have been made to any of the sets of estimates, the positions of the lines of points indicate that the LS and ME modules were found to be very similar in difficulty to the General subtests; the GA and SS modules, on the other hand, appear to have been somewhat harder than the General subtests, and the PS and TN modules slightly easier.

Departures from straight lines are very slight in all 6 graphs, though in some cases (e.g. LS, PS, TN) perhaps more noticeable than in the comparisons concerning the General subtests. Thus, following Bejar, it would again be concluded that there is little evidence to suggest that the Modular and General subtests tap different dimensions.

5.5.2.3 Subtests Treated Singly and in Combination: Comparison of Ability Estimates

The alternative method for investigating the dimensionality of data, demonstrated in Chapter 4 (Section 4.5.2.3), was also applied to the ELTS data. In this case, ability estimates were obtained using each of the subtests separately, and using the same combinations of subtests as in the previous section. It was thus possible to compare the ability estimates obtained for the same persons (a) from each General subtest treated singly and the two General subtests combined, and (b) from each Modular subtest treated singly and the same subtest combined with the two General subtests. These comparisons involved all persons in the

relevant data sets except for those scoring zero or full marks on the subtest in question.

The pairs of ability estimates are plotted in Figures 5.21 to 5.28. As was explained in Chapter 4, the points should form a straight line if the abilities involved in each case are either the same or highly correlated (though, of course, with some scatter as a result of measurement error). As far as the comparisons for the General subtests are concerned (see Figures 5.21 and 5.22), it can be seen that the majority of points cluster quite closely along a line, indicating that the two sets of ability estimates are in general highly correlated. There are, however, some rather more widely-spaced points at the extremes in each case: these are similar to some of the patterns observed in Chapter 4, and again result from a combination of (a) the large differences in ability estimates corresponding to differences of 1 raw score point at the extremes of the score range, and (b) the difference between the larger and smaller item sets in terms of the ability levels which they can differentiate.

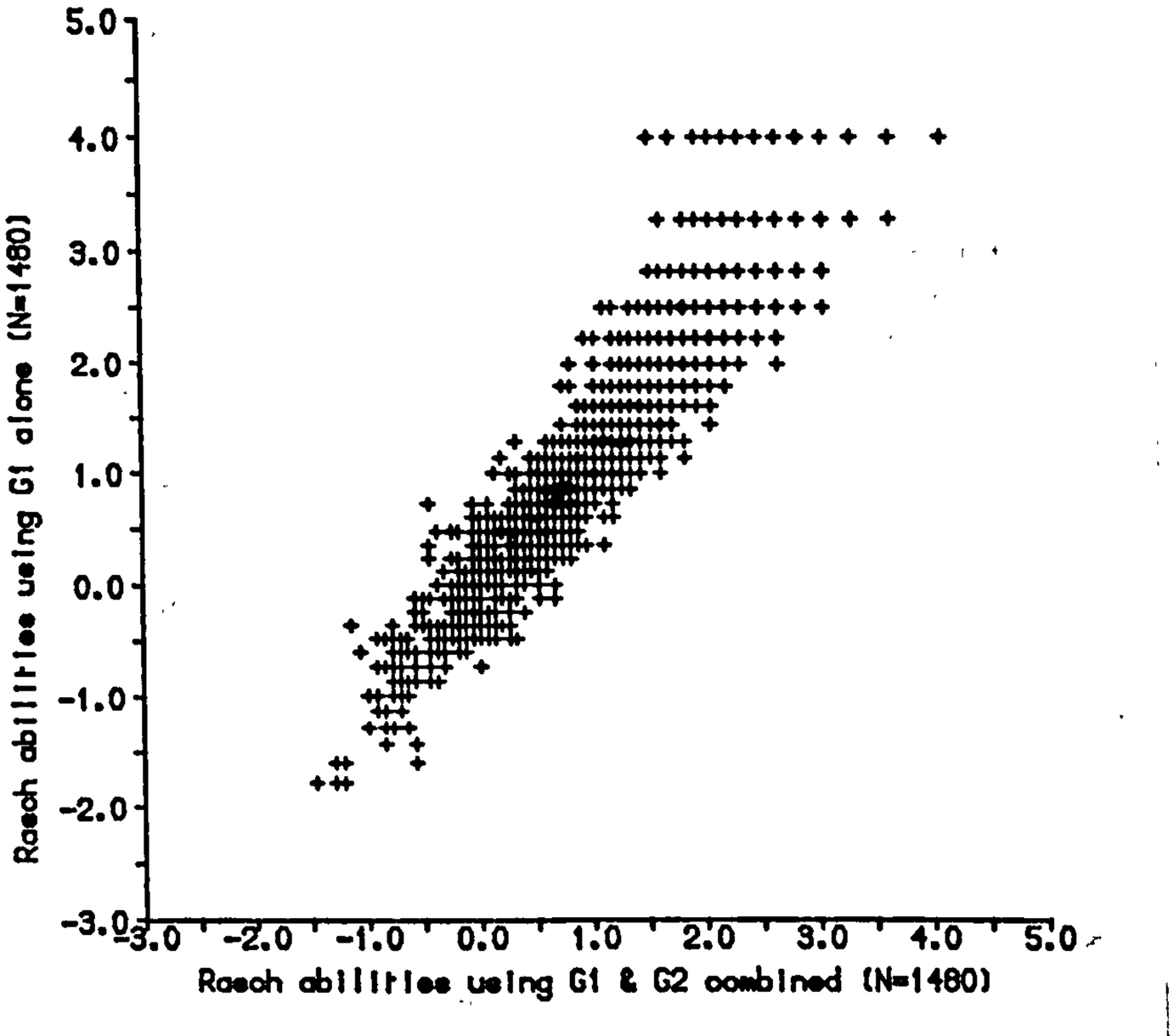


Figure 5.21 Rasch Abilities: G1 Alone vs G1 & G2 Combined

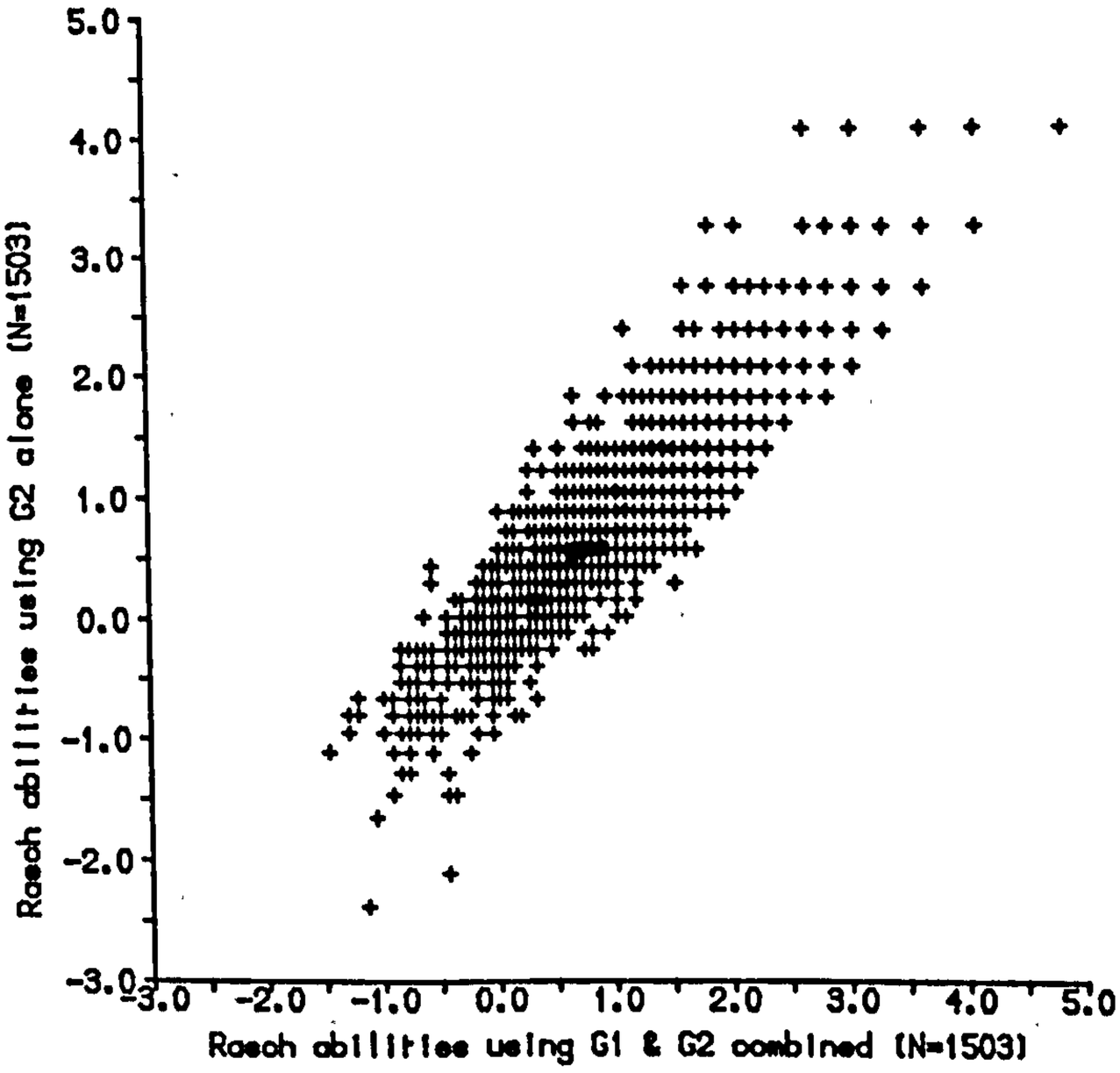


Figure 5.22 Rasch Abilities: G2 Alone vs G1 & G2 Combined

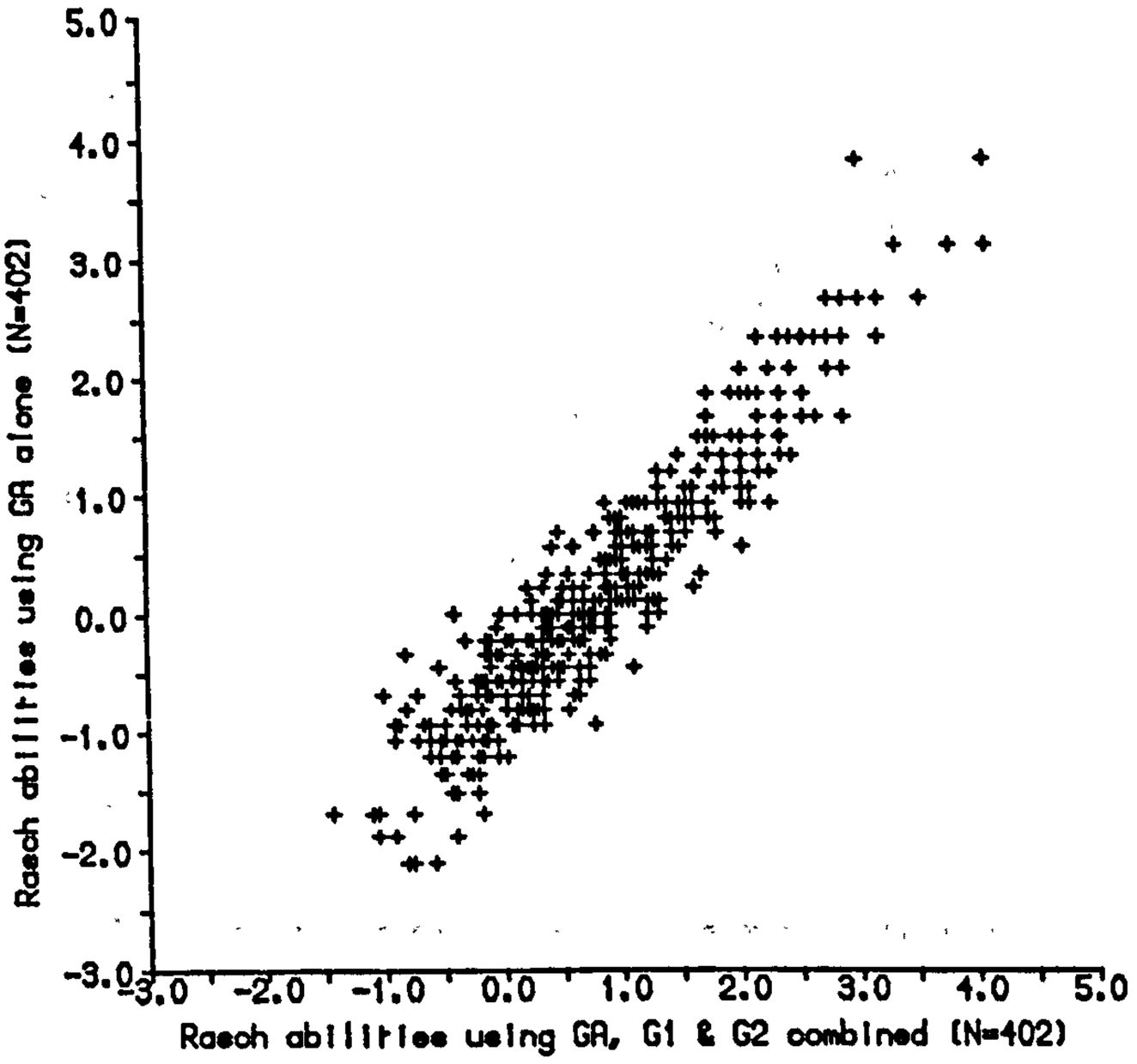


Figure 5.23 Rasch Abilities: GA Alone vs GA, G1 & G2 Combined

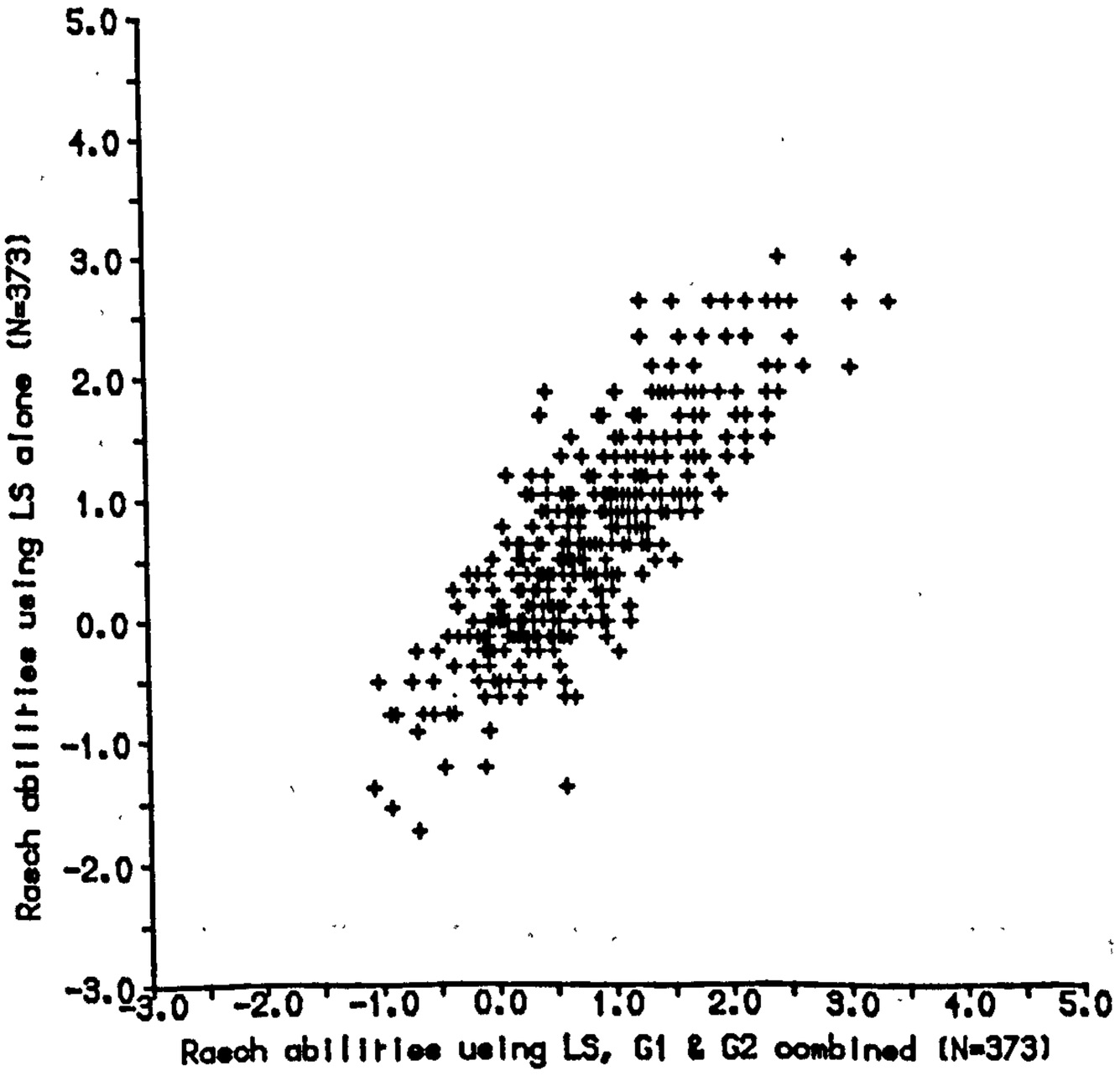


Figure 5.24 Rasch Abilities: LS Alone vs LS, G1 & G2 Combined

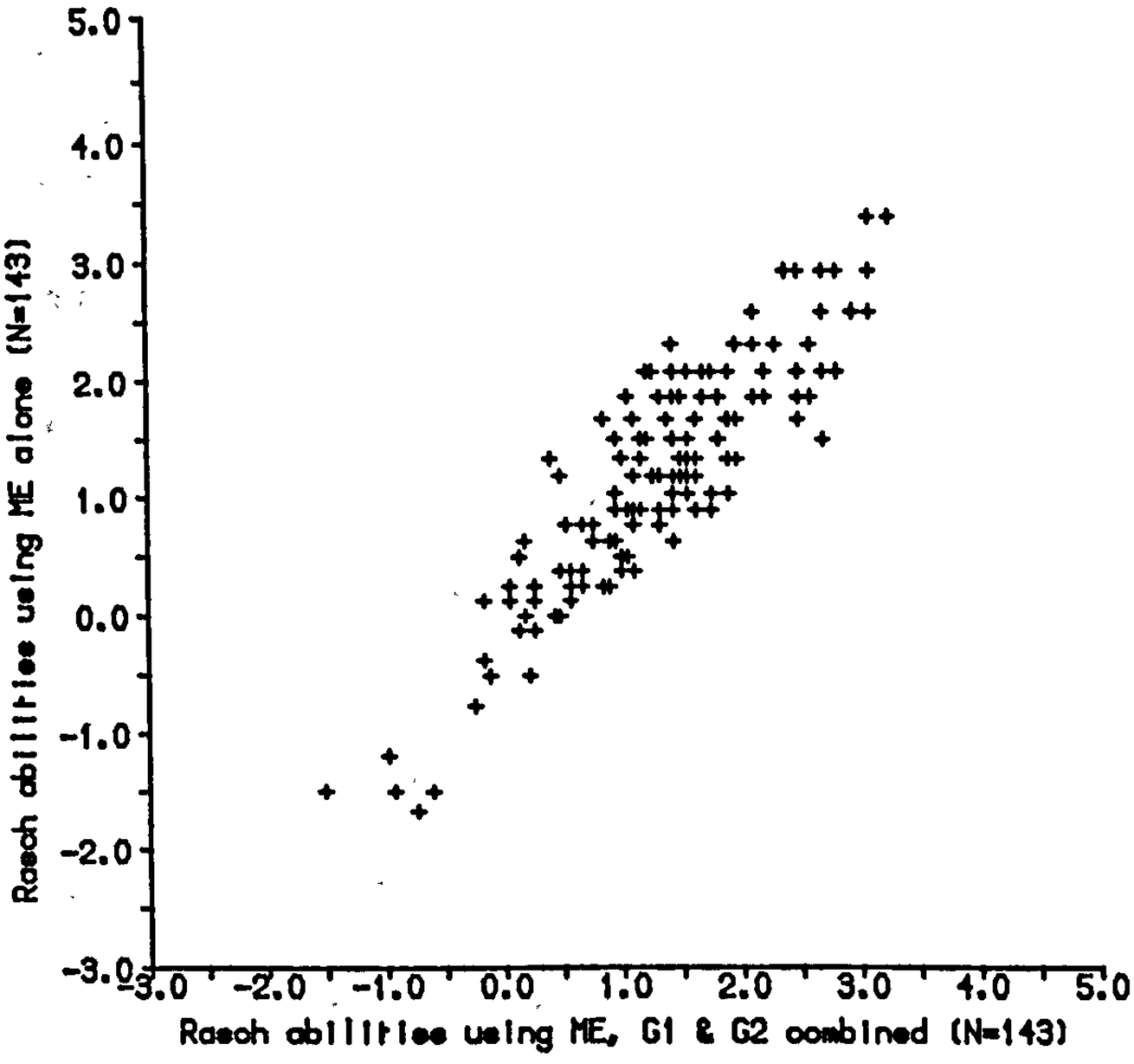


Figure 5.25 Rasch Abilities: ME Alone vs ME, G1 & G2 Combined

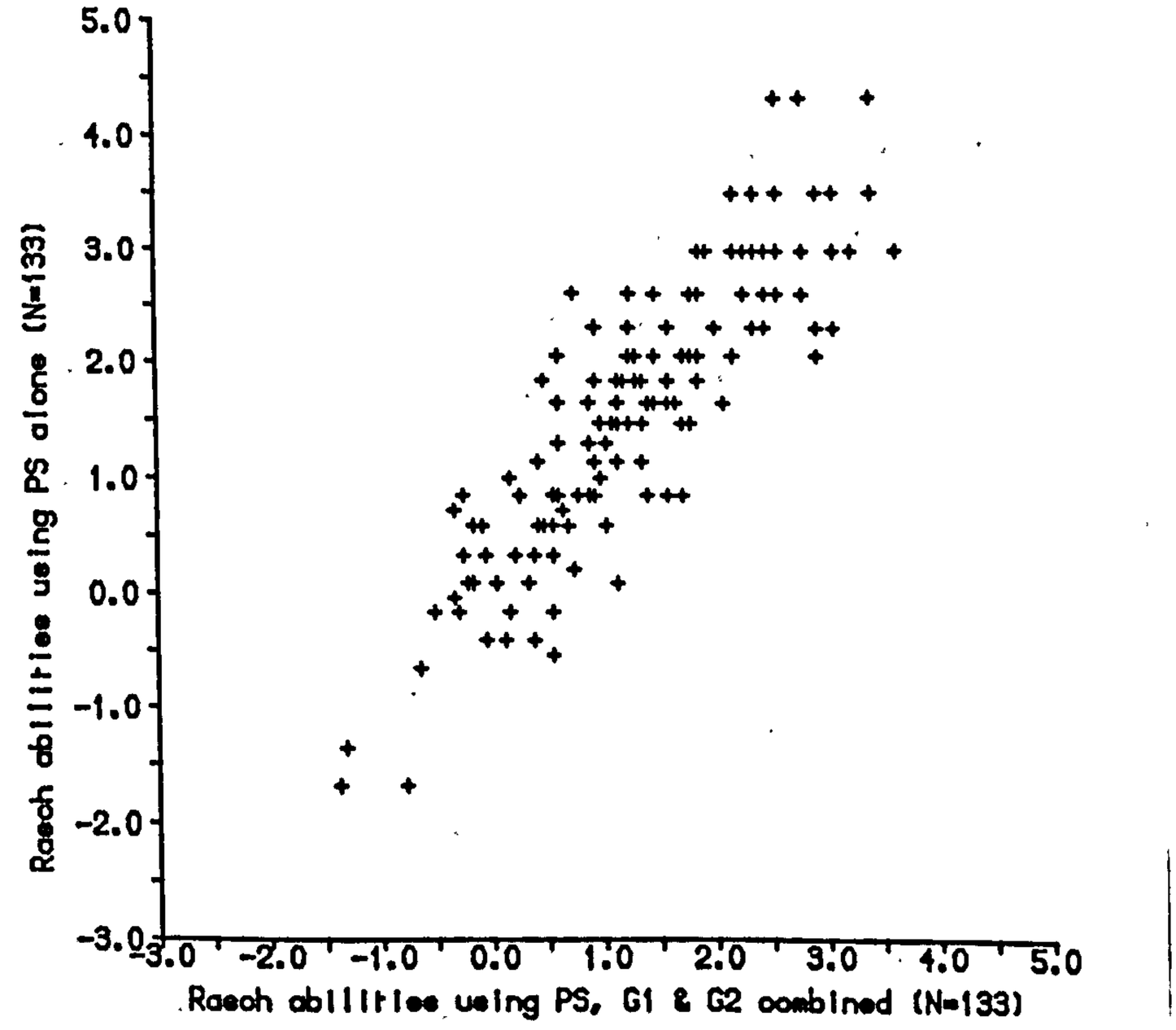


Figure 5.26 Rasch Abilities: PS Alone vs PS, G1 & G2 Combined

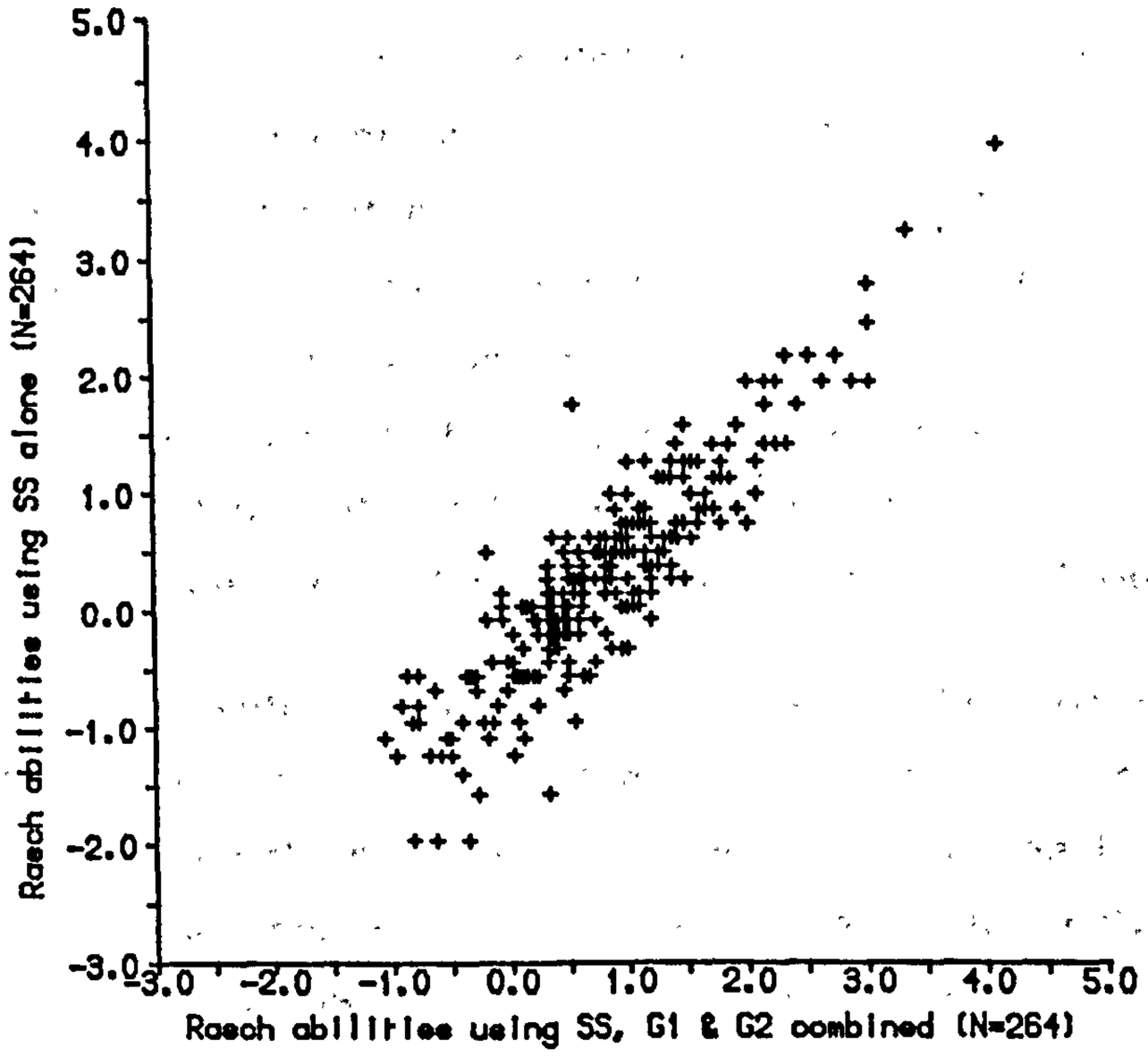


Figure 5.27 Rasch Abilities: SS Alone vs SS, G1 & G2 Combined

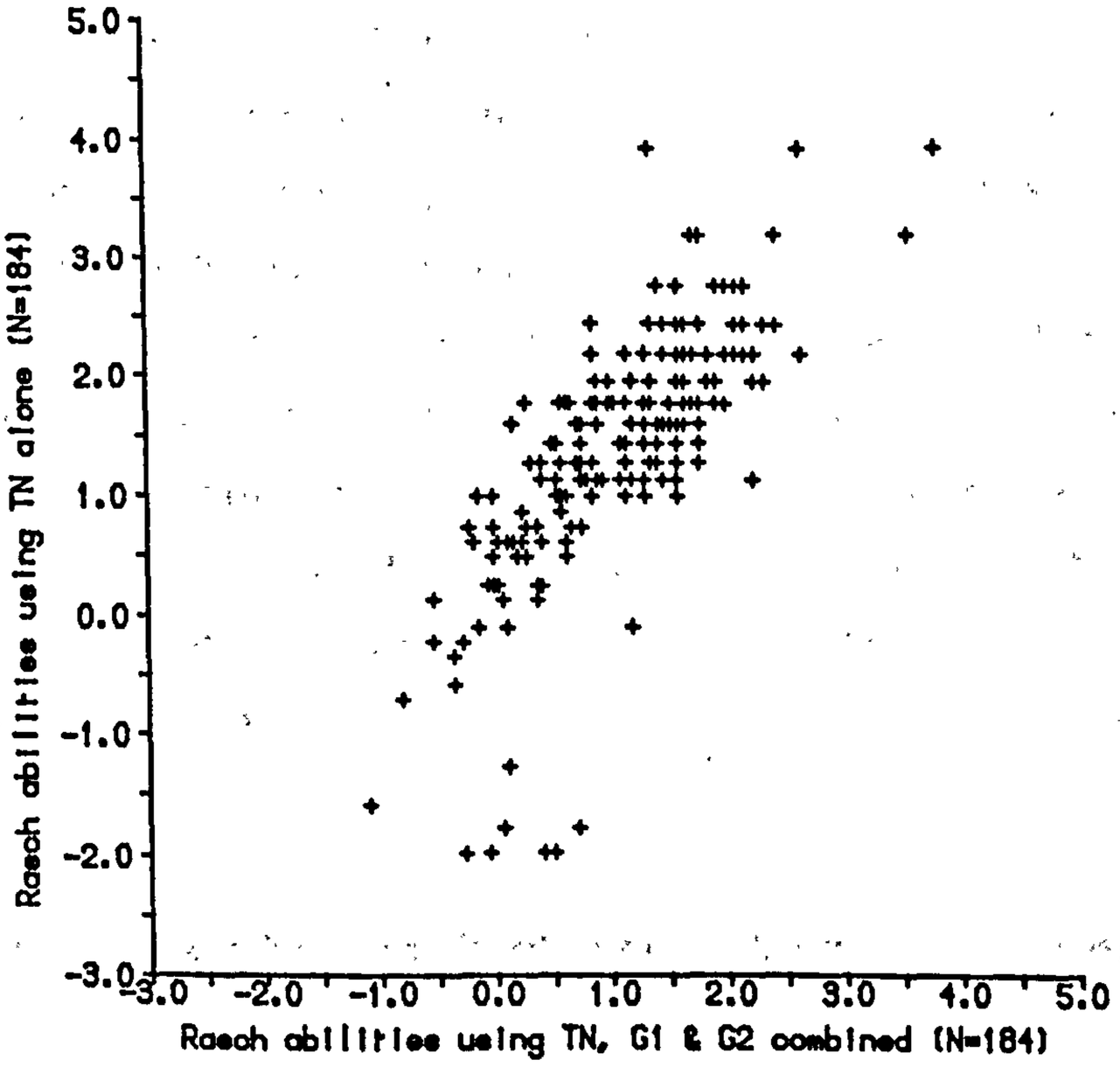


Figure 5.28 Rasch Abilities: TN Alone vs TN, G1 & G2 Combined

The comparisons involving abilities from the Modular subtests vs those from the Modular and General subtests combined (see Figures 5.23 to 5.28) also show that the sets of ability estimates are, in general, linearly related. There is, however, some variation in the degree of correspondence between the pairs of points for the different modules: those for GA, ME and SS appear, from these figures, to show the highest correspondence, while those for LS and PS appear to be more widely dispersed about their lines. The points plotted for TN show the least correspondence, particularly at the extremes. Although one might expect some inconsistency at the low extreme of the ability scale in each case, as a result of guessing, the differences noted here extend throughout the ability range.

Compared with the equivalent results for the cloze-type test (see Section 4.5.2.3), the points plotted for the ELTS subtests are in all cases more widely dispersed. It must be remembered, however, that each ELTS subtest is shorter than any of the item subsets used in the cloze-type analyses, with the result that the number of possible points on the y-axis is more restricted in each case, and error of measurement generally greater. This would account for the frequent occurrence of patterns of horizontal lines in these figures.

Comparing the results for the ELTS subtests with each other, however, there is some indication that the various modules differ in the closeness of relationship between the scores from these alone and the scores from the Modular and General subtests combined. The results would tend to suggest that some modules tap abilities other than those measured by the General parts of the test, while other modules tap abilities which are either the same as or highly correlated with those involved in the General components. The 3 modules identified here as showing least correspondence with the General subtests (LS, PS and TN) are the same as those for which very slight departures from a straight line were noted in the plots of difficulty estimates shown in the previous section. The use of ability estimates for checks of this kind can, however, be seen to provide a clearer picture of the relationships between the various test components, and to make such discrepancies as there are more easily discernible.

5.5.2.4 Subtests Treated Singly and in Combination: Comparison of Misfitting Items

If the 3 separate components of the ELTS test under consideration here (i.e. Reading, Listening and Study Skills) defined separate, uncorrelated dimensions, then one might expect at least some of the items identified as misfitting in the

combined analyses of the 3 subtests to be among those showing acceptable fit in the individual analyses.

The items for which the total fit t-statistic was greater than 2 in the combined analysis for each module are listed below. Those which were also among the misfitting items in the individual subtest analyses are marked with an asterisk. The items listed here have been grouped by the subtest to which they belong, but within these groups are arranged in order of fit, beginning with the most misfitting in each case.

G1 + G2 + GA:	G111*,	G112*,	G125*		
	G235*,	G233*,	G221,	G227*	
	GA03*,	GA05*,	GA17*,	GA13*,	GA26*
G1 + G2 + LS:	G111*,	G125*,	G109*		
	G233*,	G203,	G227*		
	LS04*,	LS16*,	LS08*,	LS14*,	LS22*
G1 + G2 + ME:	G109*				
	G228				
	ME35*				
G1 + G2 + PS:	G111*				
	G233*,	G225			
	PS06*				
G1 + G2 + SS:	SS14*,	SS22*			
G1 + G2 + TN:	G132*				
	G235*,	G209			
	TN38*,	TN13*,	TN24*		

It can be seen that each of these sets of misfitting items contains no more than one item which had not previously been identified in the individual subtest analyses (see Section 5.3.2.4 for lists of those identified for each subtest). Thus for the most part, the same items are found to show misfit both in the individual and combined analyses, which suggests that they involve some factor which is related neither to the particular subtest nor to the larger measures formed by combining the subtests.

For those which showed significant misfit only in the combined analysis, the total fit t-values in the individual subtest analyses were as follows:

G211: 0.28; G203: 1.70; G228: -0.41; G225: 0.58; G209: -1.87

It will be noted that all of the items in question belong to the Listening subtest.

As is indicated by the total fit t-values shown, 4 of these items would be considered to fit well in the separate analysis; indeed, one of them (G209) is among the 10 best-fitting items in G2. It must be remembered, of course, that the person samples used for each of the combined analyses are different, since each person answered only one Study Skills module: one might therefore expect some differences in the misfitting items on these grounds alone. It is nevertheless of interest that the discrepancies should all involve a particular subtest. However, since the items concerned come from different subsets within the Listening subtest, and since each one shows misfit in relation to only one Modular subtest (and hence to only one subgroup of candidates) it seems unlikely that the explanation for these results lies in the dimension defined by the Listening component as a whole.

5.5.3 Sample-Independence of Difficulty Estimates

The comparison of difficulty estimates for persons of different nationalities or language backgrounds was not feasible for the ELTS data, since the person sample was so diverse. This section is therefore concerned wholly with the sample-independence checks carried out on the difficulty estimates obtained for the two General subtests using the high- and low-scoring groups of 500 described in Section 5.4.1.

Following the method outlined in Chapter 4 (Section 4.5.3), 95% confidence boundaries were calculated for the item points plotted for the high- and low-scoring subgroups on each of G1 and G2. (The relevant difficulty estimates and standard errors are listed in Appendices J.8 and J.9.) The same procedure was applied to sets of difficulty estimates obtained using random groups of 500 drawn from the complete person sample, to provide a comparison with the grouping by score level.

The results of these checks are shown, for G1, in Figures 5.29 and 5.30, and, for G2, in Figures 5.31 and 5.32. It can be seen from Figure 5.29 that almost half of the plotted points for G1 fall outside the 95% confidence boundaries, indicating that the sets of estimates are not entirely sample-independent. In Figure 5.30, by contrast, only one of the points falls outside the boundaries drawn. Comparison of these two figures indicates that although the sets of estimates obtained using the high- and low-scoring subgroups are fairly consistent, the use of these particular groupings has exerted some influence on the estimates.

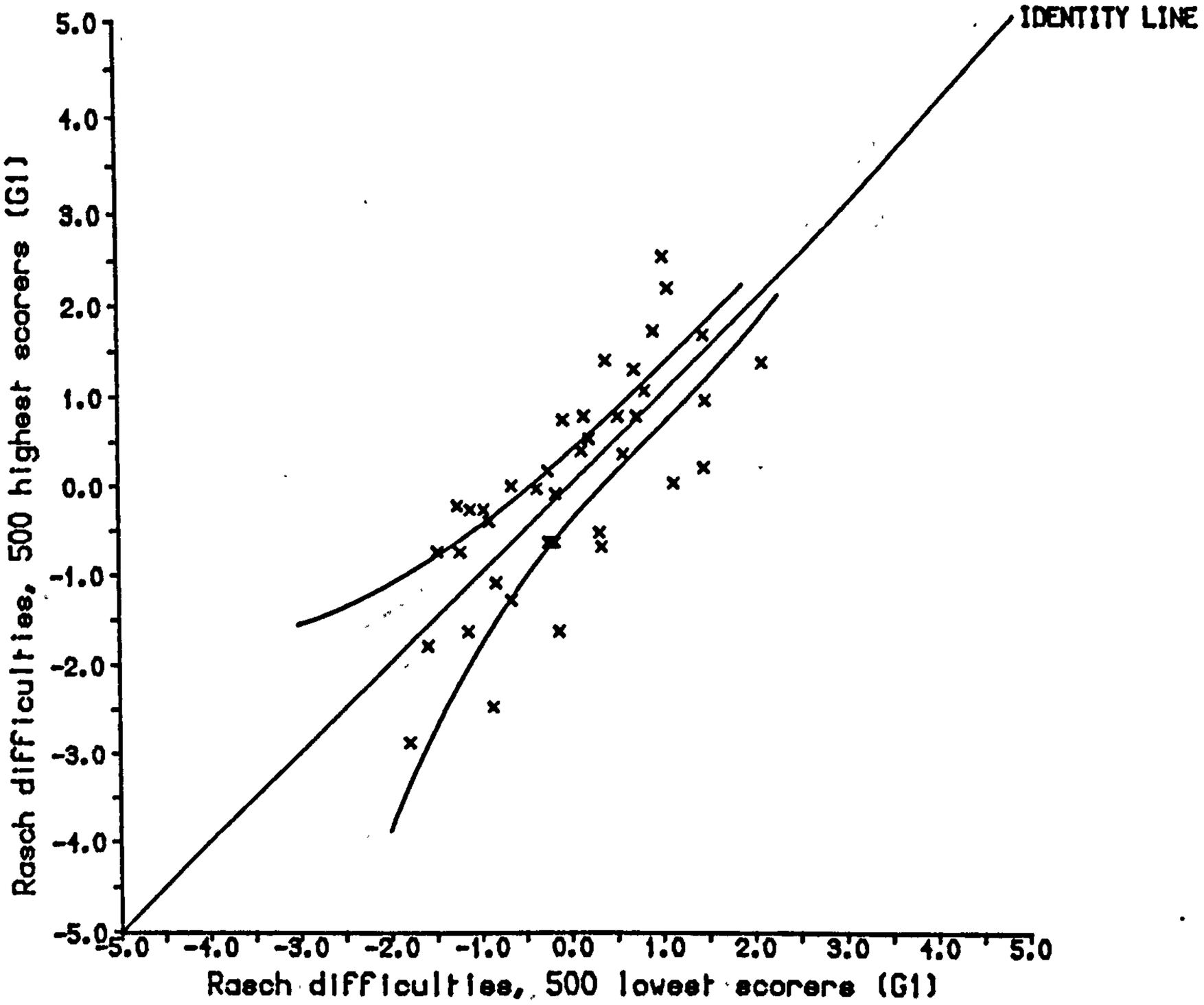


Figure 5.29 Sample-Independence Check (G1), High vs Low Scorers

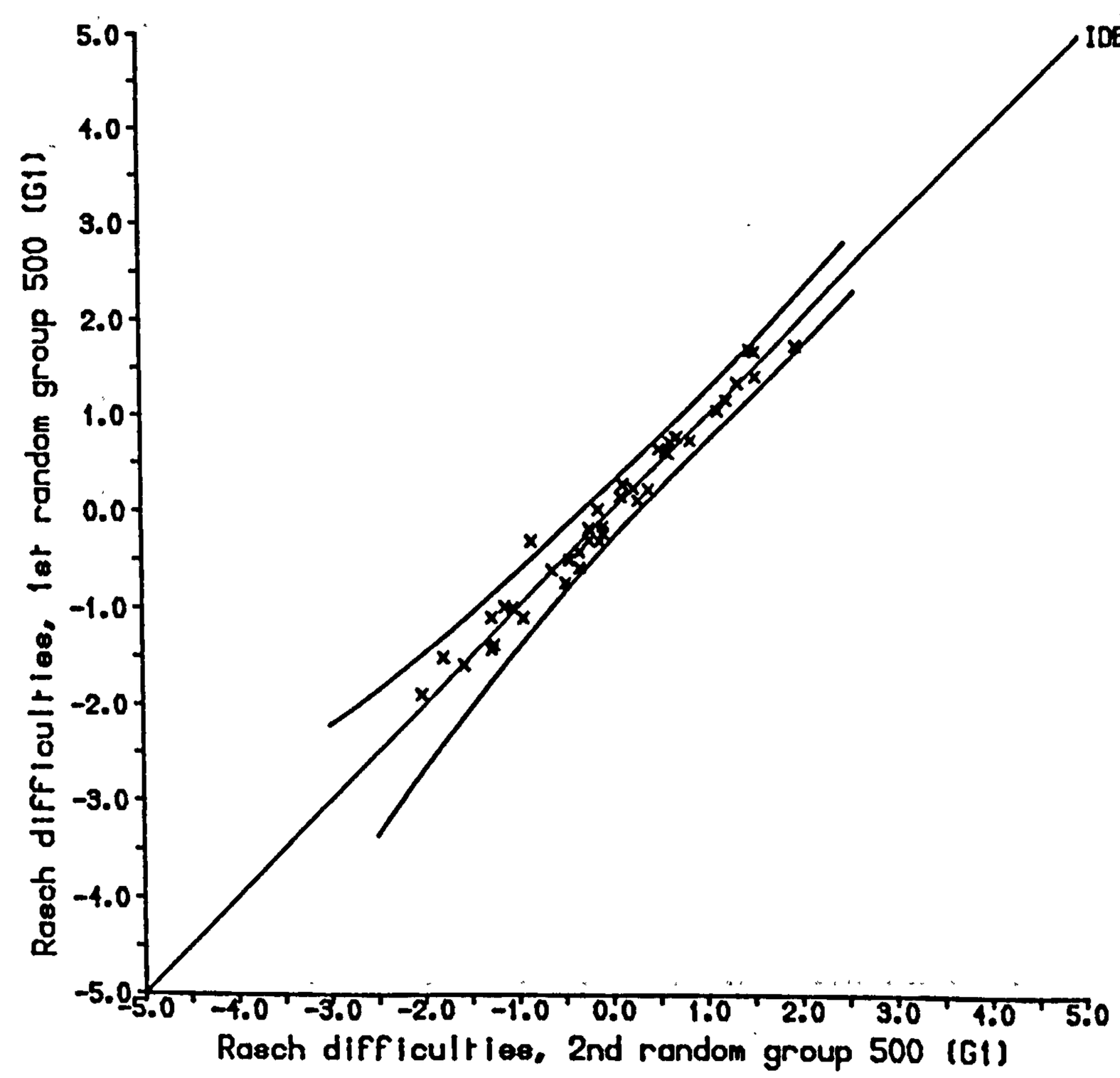


Figure 5.30 'Baseline' Plot (G1), Random Groups of 500

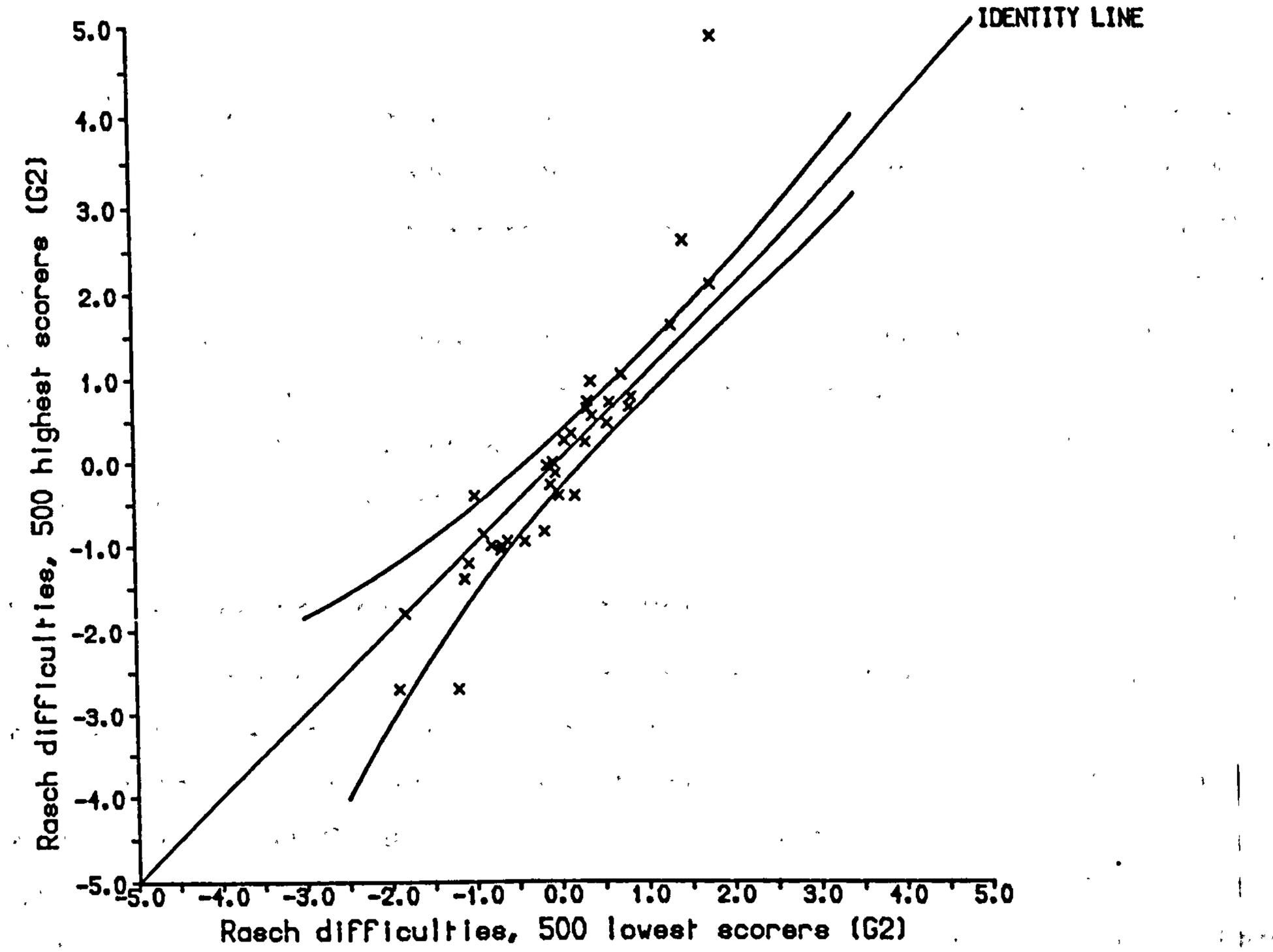


Figure 5.31 Sample-Independence Check (G2), High vs Low Scorers

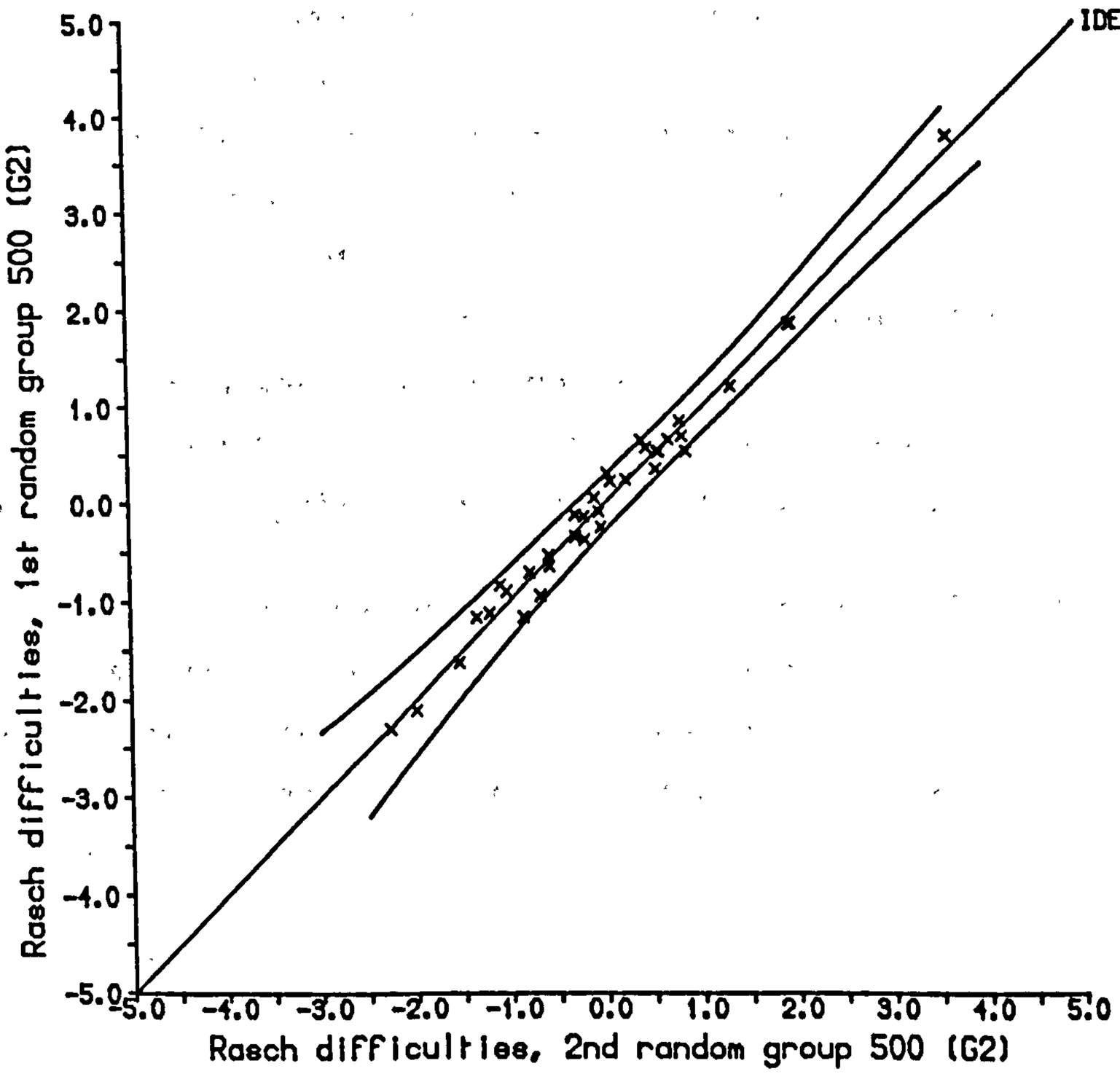


Figure 5.32 'Baseline' Plot (G2), Random Groups of 500

The results for G2 (see Figures 5.31 and 5.32) show a similar effect. However, the proportion of points falling outside the confidence limits for the high- and low-scoring groups (less than a third) is smaller than for G1, which confirms the impression that the sets of difficulty estimates calculated for G2 show greater stability. It would appear, though, that fit has still not been sufficiently good for sample-independence of estimates to be achieved. (The outlying point near the top of Figure 5.31 is that of item G227, which has been shown throughout to give rise to serious inconsistency.)

5.5.4 Test-Independence of Ability Estimates

Since the number of items in each subtest is rather small for meaningful comparison of ability estimates calculated using hard vs easy item subsets, the results reported in this section are all based on comparisons of ability estimates obtained from the General vs Modular components of the test. The checks carried out here represent extensions to those described in Section 5.2.2.3 in the sense that the element of self-correlation involved in those analyses is here removed.

Thus for all persons with scores other than zero or full marks, ability estimates were calculated separately using (a) the relevant Modular subtest, and (b) the two General subtests combined. It would have been possible to adjust the resultant sets of ability estimates in the manner described in Section 4.5.4.2, to take account of the differences between the separate and combined calibrations of the two parts of the test in terms of the difficulty estimates obtained. However, since the effect of such adjustments would simply have been to centre the plotted pairs of ability estimates around the identity line (without in any way changing their distribution), this further step was omitted here.

The results are shown, for each M1 subgroup, in Figures 5.33 to 5.38. The patterns observed are, of course, similar to those already considered in Section 5.5.2.2; however, the removal of the element of overlap in the sets of items used allows the relationship between the Modular and General subtest scores to be seen more clearly.

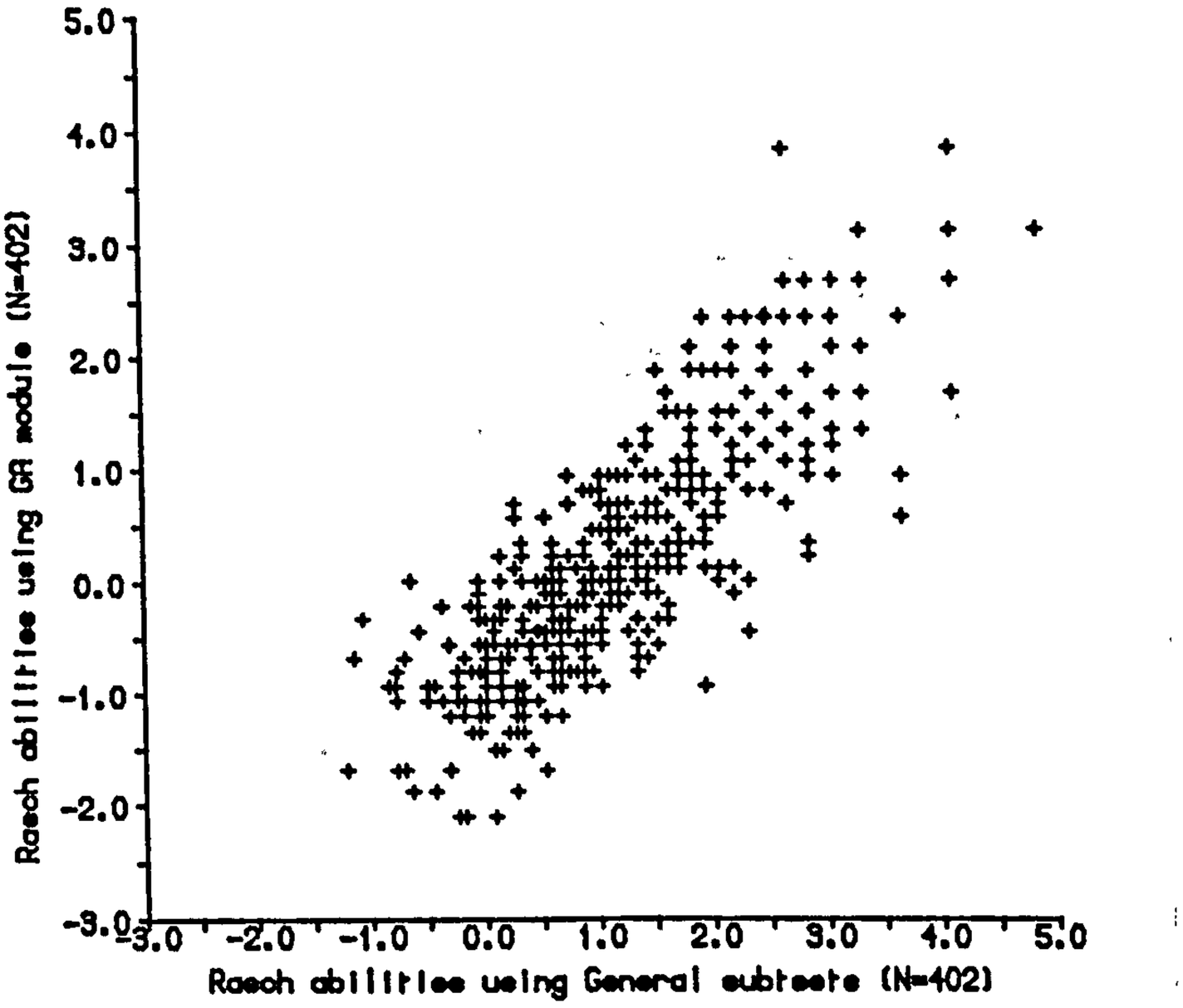


Figure 5.33 Ability Estimates, GA Module vs General Subtests

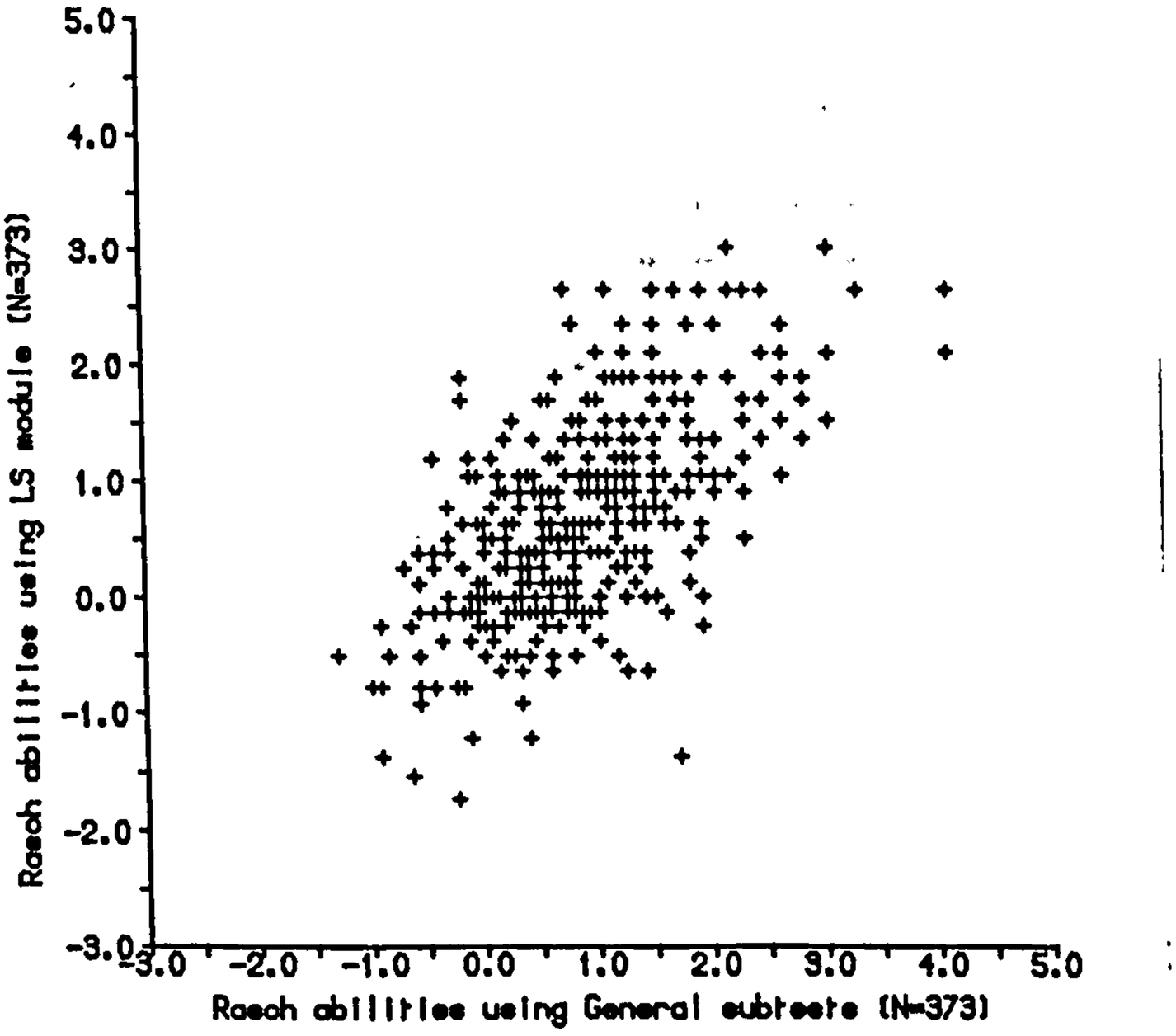


Figure 5.34 Ability Estimates, LS Module vs General Subtests

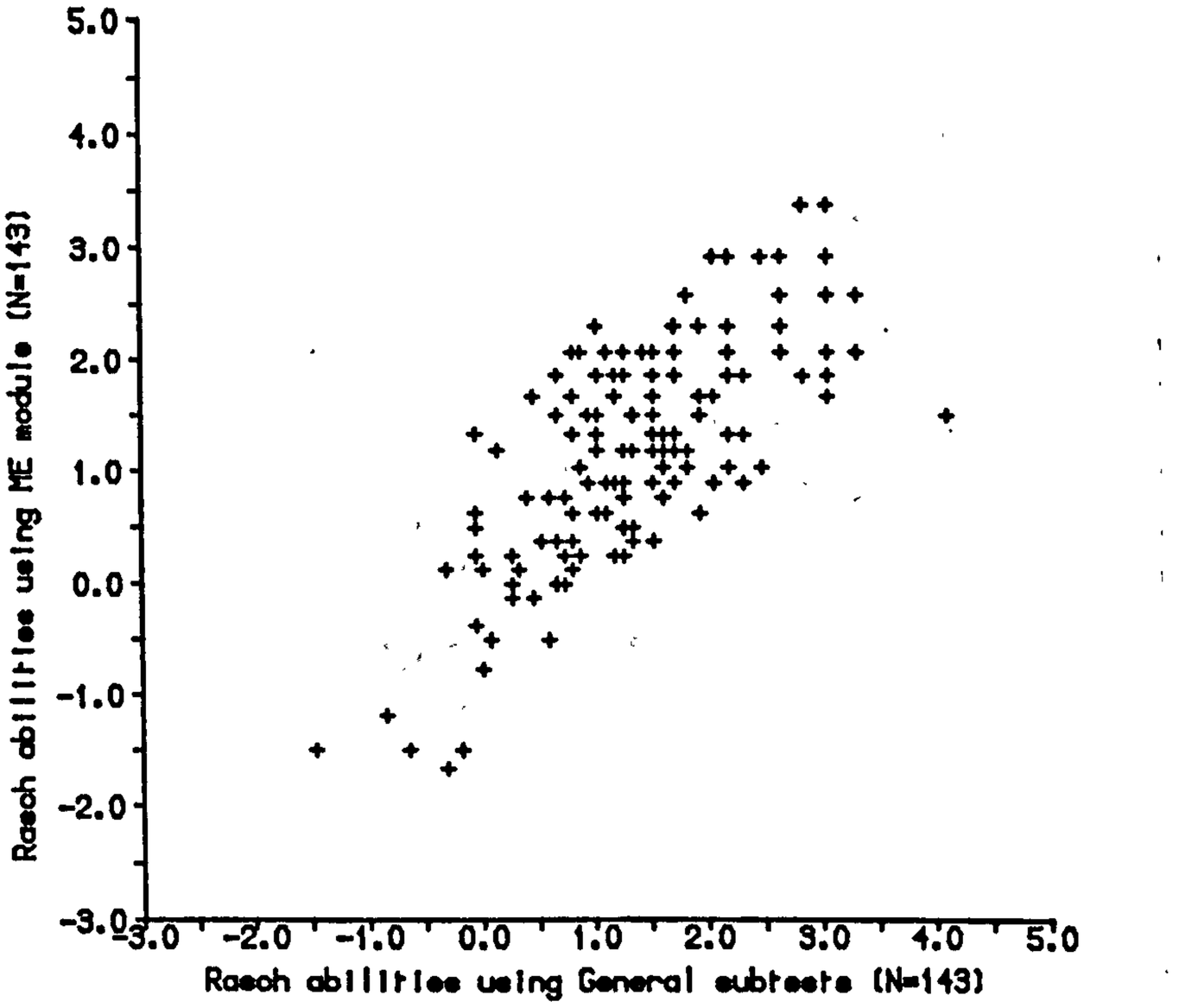


Figure 5.35 Ability Estimates, ME Module vs General Subtests

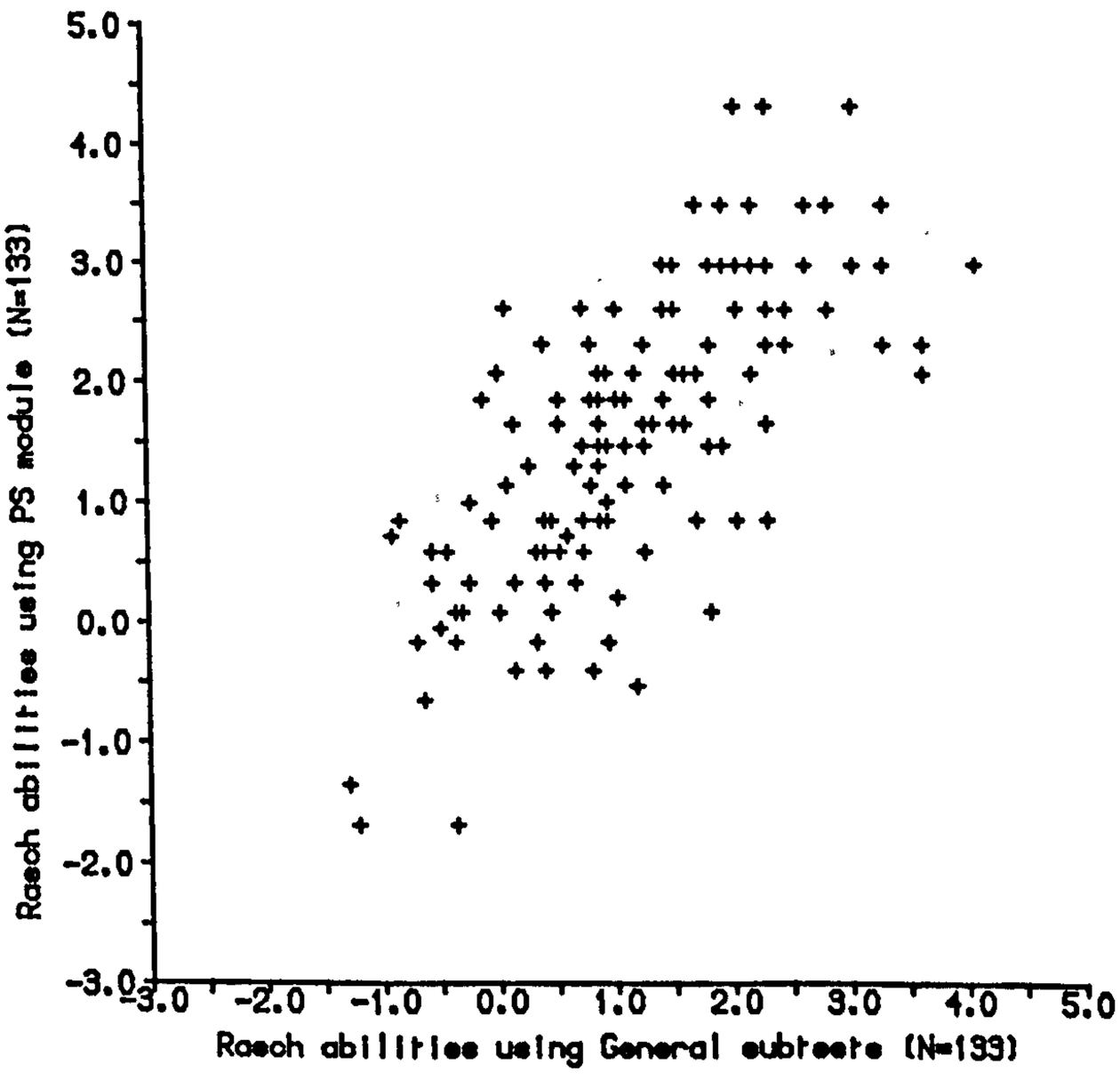


Figure 5.36 Ability Estimates, PS Module vs General Subtests

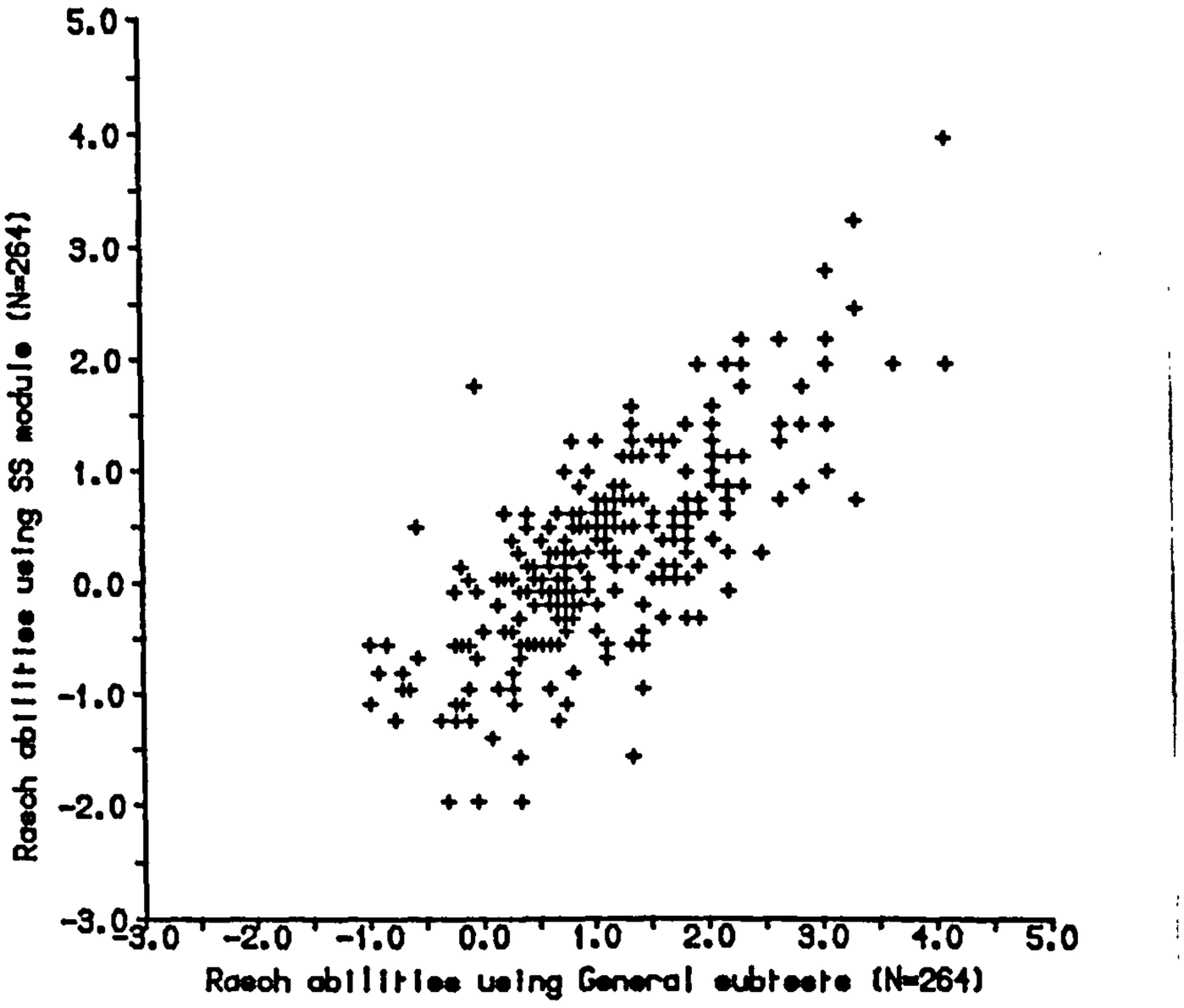


Figure 5.37 Ability Estimates, SS Module vs General Subtests

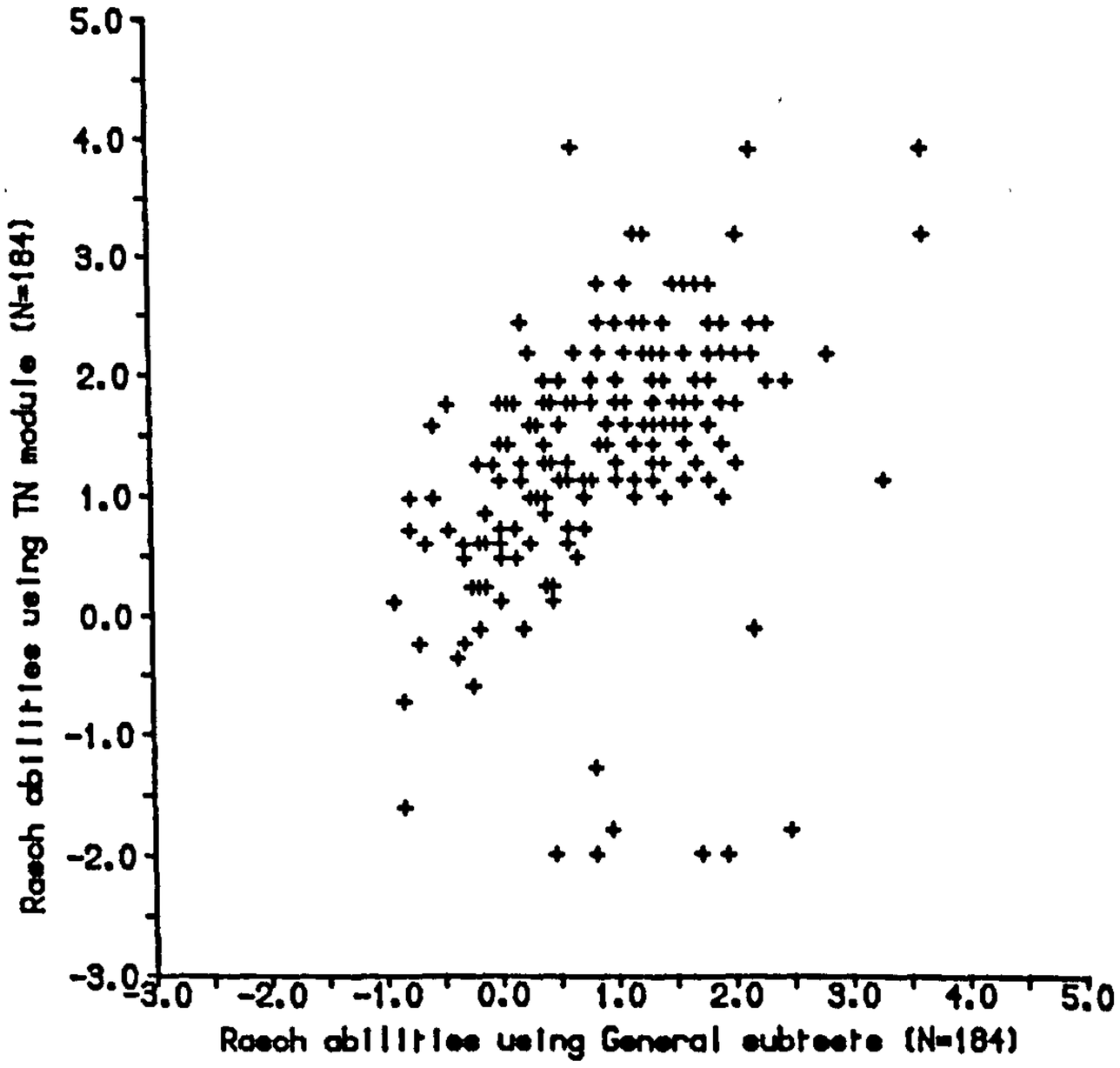


Figure 5.38 Ability Estimates, TN Module vs General Subtests

These 6 figures indicate that the extent to which the Rasch abilities depend on the particular set of items used in their estimation varies according to the Modular subtest taken. Again, it would appear that consistency between the pairs of estimates is, in general, lowest for the LS, PS and TN modules. As was suggested earlier, it is possible that these 3 Modular subtests, when combined with the items from G1 and G2, form item sets which depart more widely from unidimensionality than those formed from the General subtests combined with the GA, ME and SS modules.

5.6 Summary of Findings

In this section, the main points arising from the preceding discussion of the ELTS data analyses are summarised.

As regards the comparison of the results yielded by the traditional and Rasch analyses, the outcome was, in general, similar to that observed for the cloze-type test analyses discussed in Chapter 4: the Rasch difficulty estimates, for example, were again shown to offer greater stability across person subgroups than the facility values, and inspection of the Rasch person fit statistics and standardized residuals again provided potentially useful information which was not available using traditional procedures. One difference noted, however, was that the items identified as questionable by the traditional discrimination statistics and the item fit statistics showed greater correspondence for the ELTS data than for the cloze-type data; it was suggested that this might be due to the differences between the two in terms of the ranges of difficulty and ability spanned.

As far as the information obtained from the Rasch analyses was concerned, it was found that the patterns of person and item fit observed for the ELTS and the cloze-type data differed somewhat, no doubt as a result of the differences in the format and composition of the two tests. Some of the instances of person misfit observed in the ELTS analysis, for example, appeared to relate to the beginning of certain item subsets within a given ELTS subtest, or to chance success; neither of these applied in the case of the cloze-type data, where misfit appeared frequently to be attributable to factors relating to the marking procedure, or to idiosyncratic answering strategies on the part of some testees.

Although the item characteristic curves for the ELTS data did not, in general, depart greatly from model expectation, the results of the checks carried out here for G1 and G2 again indicated that fit was not sufficiently good for the full benefits of sample-independence of difficulty estimates to be achieved. There

were also indications that the different Modular data sets varied in their fit to the model: there was evidence of differences in the apparent effects of the time limits, and it was also suggested that the Modular subtests might differ in the extent to which they tapped the same abilities as the General subtests.

CHAPTER 6
CONCLUSIONS

Summaries of the main findings relating to the analyses of data were given at the end of Chapters 4 and 5. It is appropriate here, however, to draw together the main points which emerged, and to suggest some areas in which further investigations might prove informative, or indeed where a different approach from that used here might have been preferable.

The comparisons of the results of traditional and Rasch analyses were carried out both in order to illustrate the relationship between the two approaches (thereby demonstrating in a practical way many of the points previously set out in the theoretical background to the two approaches, presented in Chapter 2) and in order to assess their relative merits. Although the unfamiliarity of the logit scale was suggested as a possible obstacle to the acceptance of Rasch-based techniques, at least in the short term, in all other aspects the results of the comparisons pointed to the advantages offered by the Rasch approach. This was true particularly of the cloze-type data, in which traditional indices of discrimination were heavily influenced by the extreme easiness or difficulty of a number of the items: in this case the use of the item fit statistics for purposes of identifying suspicious items was clearly preferable. In the case of the ELTS subtest analyses, there was closer correspondence between the traditional and Rasch indices of item quality, as a result, it was suggested, of the closer matching of person and item levels in those data sets.

The inspection of items found to misfit (or to show low discrimination without also being of extreme easiness or difficulty) proved informative in the case of the cloze-type analyses: on the basis of this, a number of amendments to the marking scheme were suggested, since in several cases the misfit could be attributed to the fact that fairly high-level testees had thought of answers other than those specified. Indeed, it was noted that in some cases the answers given were stylistically preferable to the 'correct' answers, and it was suggested that the use of simplified reading passages for tests of this type might not be appropriate at higher proficiency levels.

The availability of the two different person samples (Malaysian and Tanzanian) allowed comparison of the two sets of item fit statistics. These were found to show only moderate correspondence, and illustrated the need to check for fit in each new application.

Since the content of particular items in the ELTS test could not be discussed, consideration of item fit was necessarily less satisfactory in that case; however, general observations were made concerning e.g. the position of misfitting items within their subtests, and the particular subset of items in which they occurred. Those who are familiar with the content of the ELTS test could no doubt suggest possible reasons for misfit in some cases, upon inspection of the items concerned.

The person fit statistics were found to offer potentially very useful information concerning the validity of test scores for individuals; it was noted that no analogous procedure existed under the traditional approach. The patterns of person misfit identified showed differences for the two different test-types, relating to differences between constructed response items and multiple choice items in terms e.g. of the possibility of correct guessing. The onset of a new passsage or item type in the ELTS subtests also appeared to lie behind some instances of person misfit; no such effect relating to the beginning of new passages in the cloze-type test was noted, however, suggesting that the item subsets in the ELTS subtests were 'more different from each other' than the passages in the cloze-type test.

The possibility of matching up persons and items on the common ability/difficulty scale was noted as a further useful feature of the Rasch approach.

The comparison of the traditional and Rasch-based indices of item difficulty showed for both tests that the latter were considerably more stable across groups differing widely in ability levels. For the two nationality groups tested on the cloze-type test, the results were less extreme (since the two groups differed less in levels), but nevertheless favoured the Rasch-based index. The additional comparisons carried out for the cloze-type test, using two further indices of difficulty, clearly demonstrated the advantage of the Rasch scale in being freed from the distribution of the original proportion-correct scores. Although less apparent in the comparison involving the Malaysian and Tanzanian groups, this effect showed clearly when the comparison involved high- and low-level groups. This might be seen to be an extreme, and hence unrealistic example; certainly, for practical purposes one would be unlikely to select samples such as these. It does, however, illustrate certain interesting characteristics of the scales.

For both the cloze-type test and the ELTS subtests, the visual presentation of the observed item characteristic curves, which had been calculated as part of the BICAL program, was found helpful in identifying between-group misfit at a glance,

and in judging the extent to which the Rasch assumption of equal discrimination appeared to hold. In the case of some of the ELTS data subsets, it would perhaps have been better to select a smaller group size for the division into ability subgroups, since in some cases the score groupings used here may have represented spurious divisions; using smaller subgroups, genuine departures from expectation could have been more easily identified.

It would also have been of interest, in view of the division of the ELTS test into separate subtests supposedly measuring different skills, to have plotted, in addition, the ICC for each item in the combined subtest calibrations. This would have been extremely uneconomical in terms of space, since each item from G1 and G2 would have appeared separately with each of the six Modular subtests; as an investigation of the dimensionality of the data, however, it would no doubt have been informative.

Guessing was not thought likely to have occurred in the cloze-type test, but was less easy to assess in relation to the ELTS test. In the latter case, intermittent high positive residuals were taken as evidence for possible chance success, though this may not have been the correct interpretation in all cases.

The ELTS subtests (with the exception of the Listening component) showed the effects of the time limit to varying degrees; the evidence examined in this regard was the steepness of the ICCs for items occurring at the end of each subtest, and the numbers of omitted items. The cloze-type test appeared to be less affected by this, however; the omissions seemed, from their positions within the test, to reflect item difficulty rather than time effects.

The comparisons of item difficulty estimates from subtest and whole-test calibrations showed little evidence of departure from unidimensionality, either for the cloze-type test or for the ELTS subtests, at least in terms of the item divisions used here. The comparisons of ability estimates obtained from the same subtest and whole-test combinations showed greater variation in all cases, though the extent to which this resulted from the greater error in the estimation of abilities using relatively few items, as compared with the estimation of difficulties using large (or fairly large) person samples, is difficult to ascertain, however. It is thus hard to summarise the extent of departure from unidimensionality indicated by these results. It is, however, tentatively suggested that the ELTS Modular subtests combined with the General subtests form tests which vary in the degree to which they measure 'the same thing'.

The feature of sample-independence of difficulty estimates did not appear to be fully achieved for the high- and low-scoring subgroups on either the cloze-type test or the G1 and G2 ELTS subtests. Stability of item difficulty estimates was, however, greater for the ELTS Listening subtest than for either of the other tests. The results of the same type of check for the Malaysian and Tanzanian groupings also indicated that model-data fit was not sufficiently good for the expected sample-independence of estimates to be achieved.

The check on the test-independence of ability estimates using the cloze-type test would have been more informative if based on less extreme item subsets. Even using the harder and easier halves of the test, though, the item subsets formed may not have been sufficiently well-matched in levels with the persons in the group in order for reasonable measures to have been achieved. Indeed, such a check is perhaps not appropriate at all for a data set which includes such a wide range of abilities and difficulties.

The checks described in the ELTS analyses as being concerned with the test-independence of ability estimates might perhaps more appropriately be viewed as further checks on the assumption of unidimensionality, since in effect they are concerned with the correspondence between scores on different parts of the same test. The degree of dispersion accounted for by measurement error, as opposed to departure from unidimensionality, could be more easily assessed by adjusting the sets of estimates to centre them around the identity line, and constructing confidence boundaries as in the checks on the sample-independence of the difficulty estimates.

Despite the comments of Raatz (1985), concerning the unsuitability of both traditional and Rasch analysis for use with cloze tests (on the grounds that local independence may not be assumed), the analyses reported here can be seen to have yielded information which could be used in improving the measures yielded by the cloze-type test. A further step in the comparison of traditional and Rasch procedures for test analysis, exemplified by Henning (1984), would be to discard the items identified as inadequate under each approach, and to reanalyse the two new sets of data, in order to see which approach brought about the greater improvement. The main problem here would, of course, be in deciding how to judge improvement; Henning compares the K-R20 reliability coefficients, but these, as was explained earlier, are influenced by a number of factors which may not reflect improvement in terms of item content.

A more interesting investigation, particularly in the light of some of the fears

expressed concerning the selection of items on statistical grounds rather than on grounds of content, might be based on a suggestion made in passing in Chapter 4. This could be applied both to the cloze-type test and the ELTS test, and would involve discarding, or modifying, those items which, on common sense grounds, seemed unlikely to result in sensible measures. Further analyses would then be carried out, to see whether these changes in fact resulted in better fit to the Rasch model.

REFERENCES

Adams, R J, Griffin, P E & Martin, L (1987) A latent trait method for measuring a dimension in second language proficiency. *Language Testing,* 4,1,9–27.

Alderson, J C (1978) A Study of the Cloze Procedure with Native and Non-native Speakers of English. PhD Thesis, University of Edinburgh.

Alderson, J C (1979) The cloze procedure and proficiency in English as a foreign language. *TESOL Quarterly,* 13,2,219–227.

Alderson, J C (1980) Native and nonnative speaker performance on cloze tests. *Language Learning,* 30,59–76.

Allen, J P B & Davies, A (Eds) (1977) *Testing and Experimental Methods. The Edinburgh Course in Applied Linguistics, Vol 4.* Oxford: Oxford University Press.

Anastasi, A (1983) Traits, states and situations: a comprehensive view. In Wainer, H & Messick, S (Eds) *Principals of Modern Psychological Measurement.* Hillsdale, N J: Lawrence Erlbaum.

Andersen, E B (1973) A goodness of fit test for the Rasch model. *Psychometrika,* 38,1,123–140.

Anderson, J, Kearney, G E & Everett, A V (1968) An evaluation of Rasch's structural model for test items. *British Journal of Mathematical and Statistical Psychology,* 21,2,231–238.

Andrich, D, De'Ath, G & Lyne, A (1982) DISLOC: A Fortran IV Program for a Rasch Model which has both a Location and a Scale Parameter for Subtests. Interim Report of a cooperative research project between the Department of Education, University of Western Australia and the State Education Department, Western Australia.

Angoff, W H (1982) Use of difficulty and discrimination indices for detecting item bias. In Berk, R.A. (Ed) *Handbook of Methods for Detecting Test Bias.* Baltimore: The John Hopkins University Press.

Bachman, L F (1982) The trait structure of cloze test scores. *TESOL Quarterly,* 16,1,61–70.

Bachman, L F (1985) Performance on cloze tests with fixed-ratio and rational deletions. *TESOL Quarterly,* 19,3,535–556.

Baker, F B (1977) Advances in item analysis. *Review of Educational Research,* 47,1,151–178.

Bejar, I I (1980) A procedure for investigating the unidimensionality of achievement tests based on item parameter estimates. *Journal of Educational Measurement,* 17,4,283–295.

Bejar, I I (1983a) *Achievement Testing. Recent Advances.* Beverly Hills: Sage

Publications.

Bejar, I I (1983b) Introduction to item response models and their assumptions. In Hambleton, R K (Ed) *Applications of Item Response Theory.* Vancouver: Educational Research Institute of British Columbia.

Birnbaum, A (1968) Some latent trait models and their use in inferring an examinee's ability. In Lord, F M & Novick, M R *Statistical Theories of Mental Test Scores.* Reading, Mass.: Addison-Wesley.

Bock, R D (1983) The discrete Bayesian. In Wainer, H & Messick, S (Eds) *Principals of Modern Psychological Measurement.* Hillsdale, N J: Lawrence Erlbaum.

Bock, R D & Lieberman, M (1970) Fitting a response model for dichotomously scored items. *Psychometrika,* 35,179-197.

Bock, R D & Wood, R (1971) Test theory. *Annual Review of Psychology,* 22, 193-224.

Brink, N E (1971) Effect of item discrimination in the Rasch model. *Proceedings of the 79th Annual Convention of the American Psychological Association.*

Brown, F G (1976) *Principles of Educational and Psychological Testing* (2nd edition). New York: Holt, Rinehart and Winston.

Brown, S (1980) *What Do They Know? A Review of Criterion-Referenced Assessment.* Edinburgh: HMSO.

Bryce, T G K (1981) Rasch-fitting. *British Educational Research Journal,* 7,2,137-153.

Carroll, J B (1968) The psychology of language testing. In Davies, A (Ed) *Language Testing Symposium.* London: Oxford University Press.

Chen, Z & Henning, G (1985) Linguistic and cultural bias in language proficiency tests. *Language Testing,* 2,2,155-163.

Choppin, B H (1976) Recent developments in item banking: a review. In De Gruijter, D N M & van der Kamp, L J Th (Eds) *Advances In Psychological and Educational Measurement.* London: John Wiley.

Criper, C & Davies, A (1986) Edinburgh ELTS Validation Project. Unpublished Project Report.

Criper, C & Dodd, W A (1984) Report on the Teaching of the English Language and its Use as a Medium in Education in Tanzania. Report for the Ministry of Education in Tanzania.

Cronbach, C J & Warrington, W G (1951) Time-limit tests: estimating their reliability and degree of speeding. *Psychometrika,* 16,167-188.

Cziko, G A (1983) Psychometric and edumetric approaches to language testing. In Oller, J W (Ed) *Issues in Language Testing Research.* Rowley, Mass.:

Newbury House.

Cziko, G A & Lin, N-H J (1984) The construction and analysis of short scales of language proficiency: classical psychometric, latent trait, and nonparametric approaches. *TESOL Quarterly,* 18,4,627-647.

Douglas, G A (1982) Conditional Inference in a generic Rasch model. In Spearritt, D (Ed) *The Improvement of Measurement in Education and Psychology.* Hawthorn, Victoria: Australian Council for Educational Research.

Ebel, R L (1972) *Essentials of Educational Measurement.* Englewood Cliffs, N J: Prentice Hall.

Goldstein, H (1979) Consequences of using the Rasch model for educational assessment. *British Educational Research Journal,* 5,2,211-220.

Goldstein, H (1980a) A rejoinder to Preece. *British Educational Research Journal,* 6,2,211-212.

Goldstein, H (1980b) Dimensionality, bias, independence and measurement scale problems in latent trait test score models. *British Journal of Mathematical and Statistical Psychology,* 33,234-246.

Goldstein, H (1981) Limitations of the Rasch model for educational assessment. In Lacey, C. & Lawton, D. (Eds) *Issues in Evaluation and Accountability.* London: Methuen.

Goldstein, H & Blinkhorn, S (1977) Monitoring educational standards - an inappropriate model. *Bulletin of the British Psychological Society,* 30,309-311.

Goldstein, H & Blinkhorn, S (1982) The Rasch model still does not fit. *British Educational Research Journal,* 8,2,167-170.

Guilford, J P & Fruchter, B (1978) *Fundamental Statistics in Psychology and Education* (6th edition). New York: McGraw-Hill.

Gulliksen, H (1950) *Theory of Mental Tests.* New York: John Wiley.

Gustafsson, J-E (1977) The Rasch model for dichotomous items: theory, applications and a computer program. Report No 63, Institute of Education, University of Goteborg, Sweden.

Gustafsson, J-E (1980) Testing and obtaining fit of data to the Rasch model. *British Journal of Mathematical and Statistical Psychology,* 33,205-233.

Gustafsson, J-E (Jan. 1981) *PML* version 3.1 Institute of Education, University of Goteborg.

Hambleton, R K (1979) Latent trait models and their applictions. In Traub, R (Ed) *New Directions for Testing and Measurement. Methodological Developments.* San Francisco: Jossey Bass.

Hambleton, R K (1980) Latent ability scales: interpretations and uses. In Mayo, S

T (Ed) *Interpreting Test Performance. New Directions for Testing and Measurement No.6.* San Francisco: Jossey-Bass.

Hambleton, R K & Cook, L L (1977) Latent trait models and their use in the analysis of educational test data. *Journal of Educational Measurement,* 14,2,75-96.

Hambleton, R K & de Gruijter, D N M (1983) Application of item response models to criterion-referenced test item selection. *Journal of Educational Measurement,* 20,4,355-367.

Hambleton, R K & van der Linden, W J (1982) Advances in item response theory and applications: an introduction. *Applied Psychological Measurement,* 6,4,373-378.

Hambleton, R K & Murray, L N (1983) Some goodness of fit investigations for item response models. In Hambleton, R K (Ed) *Applications of Item Response Theory.* Vancouver: Educational Research Institute of British Columbia.

Hambleton, R K & Swaminathan, H (1985) *Item response theory.* Boston: Kluwer-Nijhoff.

Hambleton, R K, Swaminathan, H, Cook, L L, Eignor, D R & Gifford, J A (1978) Developments in latent trait theory: models, technical issues, and applications. *Review of Educational Research,* 48,4,467-510.

Harris, D P (1969) *Testing English as a Second Language.* New York: McGraw Hill.

Heaton, J B (1975) *Writing English Language Tests.* London: Longman.

Henning, G (1984) Advantages of latent trait measurement in language testing. *Language Testing,* 1,2,123-133.

Henning, G (1986) Item banking via dBASE II: the UCLA ESL Proficiency Examination experience. In Stansfield, C W (Ed) *Technology and Language Testing.* Washington, DC: TESOL.

Henning, G (1987) *A Guide to Language Testing.* Cambridge, Mass: Newbury House.

Henning, G, Hudson, T & Turner, J (1985) Item response theory and the assumption of unidimensionality for language tests. *Language Testing,* 2,2,141-154.

Horst, P (1953) Correcting the Kuder-Richardson reliability for disperson of item difficulties. *Psychological Bulletin,* 50,371-374.

Hulin, C L, Drasgow, F & Komocar, J (1982) Applications of item response theory to analysis of attitude scale translations. *Journal of Applied Psychology,* 67,818-825.

Hulin, C L, Drasgow, F & Parsons, C K (1983) *Item Response Theory.* Homewood,

Illinois: Dow-Jones Irwin.

de Jong, J H A L (1983) Focusing in on a latent trait: an attempt at construct validation by means of the Rasch model. In van Weeren, J (Ed) *Practice and Problems in Language Testing 5. Non-Classical Test-Theory Final Examinations in Secondary Schools.* Arnhem: CITO, Dutch National Institute for Educational Measurement.

de Jong, J H A L (1984a) Testing foreign language listening comprehension. *Language Testing,* 1,1,97-100.

de Jong, J H A L (1984b) Listening, a single trait in first and second language learning. In *Toegepaste Taal Welenschap in Artikelen No.20.* Amsterdam: VII Boekhandel.

de Jong, J H A L (1986a) Item selection from pretests in mixed ability groups. In Stansfield, C W (Ed) *Technology and Language Testing.* Washington, DC: TESOL.

de Jong, J H A L (1986b) Achievement tests and national standards. In *Studies in Educational Evaluation,* 12,295-304.

Krzanowski, W J & Woods, A J (1984) Statistical aspects of reliability in language testing. *Language Testing,* 1,1,1-20.

Lee, Y P (1985) Investigating the validity of the cloze score. In Lee, Y P, Fok, A C Y Y, Lord R & Low, G (Eds) *New Directions in Language Testing.* Oxford: Pergamon Press.

Levine, M V & Rubin, D B (1979) Measuring the appropriateness of multiple-choice test scores. *Journal of Educational Statistics,* 4,4,269-290.

Linn, R L & Werts, C E (1979) Covariance structures and their analysis. In Traub, R (Ed) *New Directions for Testing and Measurement No. 4. Methodological Developments.* San Francisco: Jossey-Bass.

Loevinger, J (1965) Person and population as psychometric concepts. *Psychological Review,* 72,143-155.

Lord, F M (1952) A Theory of Test Scores. Psychometric Monograph No 7, Psychometric Society.

Lord, F M (1968) An analysis of the Verbal Scholastic Aptitude Test using Birnbaum's three-parameter logistic model. *Educational and Psychological Measurement,* 28,989-1020.

Lord, F M (1970) Item characteristic curves estimated without knowledge of their mathematical form - a confrontation of Birnbaum's logistic model. *Psychometrika,* 35,1,43-50.

Lord, F M (1974a) Individualized testing and item characteristic curve theory. In Krantz, D H, Atkinson, R C, Luce, R D & Suppes, P (Eds) *Measurement,*

Psychophysics, and Neural Information Processing. Contemporary Developments in Mathemetical Psychology Vol.II. San Francisco: W H Freeman.

Lord, F M (1974b) Estimation of latent ability and item parameters when there are omitted responses. *Psychometrika,* 39,2,247–264.

Lord, F M (1980) *Applications of Item Response Theory to Practical Testing Problems.* Hillsdale, NJ: Lawrence Erlbaum.

Lord, F M (1983) Small N justifies Rasch model. In Weiss, D J (Ed) *New Horizons in Testing.* New York: Academic Press.

Lord, F M & Novick, M R (1968) *Statistical Theories of Mental Test Scores.* Reading, Mass: Addison–Wesley.

Lumsden, J (1976) Test theory. *Annual Review of Psychology,* 27,251–280.

Madsen, H S & Larson, J W (1986) Computerized Rasch analysis of item bias in ESL tests. In Stansfield, C W (Ed) *Technology and Language Testing.* Washington, DC: TESOL.

Mead, R (1976) Assessment of Fit of Data to the Rasch Model through Analysis of Residuals. Doctoral Dissertation, University of Chicago.

Morrow, K (1979) Communicative language testing: revolution or evolution? In Brumfit, C J & Johnson, K (Eds) *The Communicative Approach to Language Teaching.* Oxford: Oxford University Press.

Munby, J (1978) *Communicative Syllabus Design.* Cambridge: Cambridge University Press.

Nitko, A J (1983) *Educational Tests and Measurement. An Introduction.* New York: Harcourt Brace Jovanovich.

Nunnally, J C (1978) *Psychometric Theory* (2nd Edition). New York: McGraw Hill.

Oller, J W (1979) *Language Tests at School.* London: Longman.

Panchapakesan, N (1969) The Simple Logistic Model and Mental Measurement. Doctoral Dissertation, University of Chicago.

Payne, D A & McMorris, R F (1975) *Educational and Psychological Measurement* (2nd edition). Morristown, NJ: General Learning Press.

Perkins, K & Miller, L D (1984) Comparative analyses of English as a Second Language reading comprehension data: classical test theory and latent trait measurement. *Language Testing,* 1,1,21–32.

Pollitt, A (1979) Item banking. In *Issues in Educational Assessment.* SED Occasional Papers, London: HMSO.

Pollitt, A (1981) Guest Lecture delivered in Dept. of Linguistics, University of Edinburgh.

Pollitt, A (1985) Statistics for Criterion Referenced Testing. Paper presented at

British Council, London.

Pollitt, A & Hutchinson, C (1987) Calibrating graded assessments. Rasch partial credit analysis of performance in writing. *Language Testing,* 4,1,72–92.

Preece, P F W (1980) On rashly rejecting Rasch: a response to Goldstein. *British Educational Research Journal,* 6,2,209–211.

Raatz, U (1985) Better theory for better tests? *Language Testing,* 2,1,61–75.

Rasch, G (1960) *Probabilistic Models for some Intelligence and Attainment Tests.* Copenhagen: Danmarks Paedagogiske Institut. Expanded edition (1980), Chicago: University of Chicago Press.

Rentz, R R & Bashaw, W L (1977) The National Reference Scale for Reading: an application of the Rasch model. *Journal of Educational Measurement,* 14,2,161–179.

Rudner, L M (1983) A closer look at latent trait parameter invariance. *Educational and Psychological Measurement,* 43,4,951–955.

Samejima, F (1977) A use of the information function in tailored testing. *Applied Psychological Measurement,* 1,2,233–247.

Samejima, F (1983) Some methods and approaches of estimating the operating characteristics of discrete item responses. In Wainer, H & Messick, S (Eds) *Principals of Modern Psychological Measurement.* Hillsdale, NJ: Lawrence Erlbaum.

Spurling, S (1987) Questioning the use of the Bejar method to determine unidimensionality. *Language Testing* (Correspondence), 4,1,93–95.

Stanley, J C (1972) Variability within individuals. In Bracht, G H, Hopkins, K D & Stanley, J C (Eds) *Perspectives in Educational and Psychological Measurement.* Englewood Cliffs, NJ: Prentice Hall.

Stansfield, C W (Ed) (1986) *Technology and Language Testing.* Washington, DC: TESOL.

Stern, H H (1983) *Fundamental Concepts of Language Teaching.* Oxford: Oxford University Press.

Stocking, M (1985) Introduction to item response theory (IRT). Seminar held at ETS, Princeton, NJ.

Subkoviak, M J & Baker, F B (1977) Test theory. *Review of Research in Education,* 5,275–317.

Swaminathan, H (1983) Parameter estimation in item response models. In Hambleton, R K (Ed) *Applications of Item Response Theory.* Vancouver: Educational Research Institute of British Columbia.

Tall, G (1981) The possible dangers of applying the Rasch model to school examinations and standardized tests. In Lacey, C. & Lawton, D. (Eds) *Issues*

in Evaluation and Accountability. London: Methuen.

Theunissen, T J J M (1987) Text banking and test design. *Language Testing,* 4,1,1–8.

Thorndike, R L (1982a) Educational measurement – theory and practice. In Spearitt, D (Ed) *The Improvement of Measurement in Education and Psychology.* Hawthorn, Victoria: The Australian Council for Educational Research.

Thorndike, R L (1982b) *Applied Psychometrics.* Boston: Houghton Mifflin.

Thorndike, R L & Hagen, E P (1977) *Measurement and Evaluation in Psychology and Education* (4th Edition). New York: John Wiley.

Traub, R E & Wolfe, R G (1981) Latent trait theories and the assessment of educational achievement. *Review of Research in Education,* 9, 377–435.

Tyler, L E & Walsh, W B (1979) *Tests and Measurements* (3rd edition). Englewood Cliffs, NJ: Prentice-Hall.

University of Cambridge Local Examinations Syndicate & The British Council (undated) *English Language Testing Service. An Introduction.*

Wainer, H (1983) Frederic M Lord: a biographical sketch. In Wainer, H & Messick, S (Eds) *Principals of Modern Psychological Measurement.* Hillsdale, NJ : Lawrence Erlbaum Associates.

Waller, M J (1981) A procedure for comparing logistic latent trait models. *Journal of Educational Measurement,* 18,2,119–125.

Weiss, D J (1982) Improving measurement quality and efficiency with adaptive testing. *Applied Psychological Measurement,* 6,4,473–492.

Weiss, D J (1983) *New Horizons in Testing.* New York: Academic Press.

Weiss, D J & Davison, M L (1981) Test theory and methods. *Annual Review of Psychology,* 32,629–658.

Whitely, S E (1977) Models, meanings and misunderstandings: some issues in applying Rasch's theory. *Journal of Educational Measurement,* 14,163–178.

Whitely, S E & Dawis, R V (1976) The influence of test context on item difficulty. *Educational and Psychological Measurement,* 36,329–337.

Willmott, A S & Fowles, D E (1974) *The Objective Interpretation of Test Performance.* Slough: National Foundation for Educational Research.

Wingersky, M S; Barton, M A & Lord, F M (1982) *LOGIST* User's Guide. Princeton, NJ: Educational Testing Service.

van den Wollenberg, A L (1980) On the Wright-Panchapakesan Goodness of Fit Test for the Rasch Model. Dept. of Math. Psychology, University of Nijmegen, The Netherlands.

Wood, R (1976) Trait measurement and item banks. In De Gruijter, D N M & van

der Kamp, L J Th (Eds) *Advances in Psychological and Educational Measurement.* London: John Wiley.

Wood, R (1978) Fitting the Rasch model - a heady tale. *British Journal of Mathematical and Statistical Psychology,* 31,27-32.

Wood, R & Skurnik, L S (1969) *Item Banking.* Slough: National Foundation for Educational Research in England and Wales.

Woods, A & Baker, R (1985) Item response theory. *Language Testing,* 2,2,117-140.

Wright, B D (1968) Sample-free test calibration and person measurement. In *Proceedings of the 1967 Invitational Conference on Testing Problems.* Princeton, NJ: Educational Testing Service.

Wright, B D (1977a) Solving measurement problems with the Rasch model. *Journal of Educational Measurement,* 14,2,97-116.

Wright, B D (1977b) Misunderstanding the Rasch model. *Journal of Educational Measurement,* 14,3,219-225.

* **Wright, B D & Douglas, G A** (1977) Best procedures for sample-free item analysis. *Applied Psychological Measurement,* 1,2,281-295.

Wright, B D & Linacre, J M (1984) Microscale Manual for Microscale Version 1.2. Westport, Connecticut: Mediax Interactive Technologies.

Wright, B D & Masters, G N (1982) *Rating Scale Analysis.* Chicago: MESA Press.

Wright, B D & Mead, R J (1975) *CALFIT* Research Memorandum No.18. Dept. of Education, University of Chicago.

Wright, B D, Mead, R J & Bell, S R (1980) BICAL: Calibrating Items with the Rasch Model. Research Memorandum No.23C, Statistical Laboratory, Department of Education, University of Chicago.

Wright, B D; Mead, R J & Draba, R E (1976) Detecting and Correcting Test Item Bias with a Logistic Response Model. Research Memorandum No.22, Statistical Laboratory, Department of Education, University of Chicago.

Wright, B D & Panchapakesan, N (1969) A procedure for sample-free item analysis. *Educational and Psychological Measurement,* 29,23-48.

Wright, B D & Stone, M H (1979) *Best Test Design.* Chicago: MESA Press.

Yen, W M (1980) The extent, causes and importance of context effects on item parameters for two latent trait models. *Journal of Educational Measurement,* 17,4,297-311.

ADDENDA TO REFERENCES

Choppin, B H (1981) Educational measurement and the item bank model. In Lacey, C & Lawton, D (Eds) *Issues in Evaluation and Accountability.* London: Methuen.

Cohen, L (1979) Approximate expressions for parameter estimates in the Rasch model. *British Journal of Mathematical and Statistical Psychology*,32,113-120.

Griffin, P E (1985) The use of latent trait models in the calibration of tests of spoken language in large-scale selection-placement programs. In Lee, Y P, Fok, A C Y Y, Lord, R & Low, G (Eds) *New Directions in Language Testing.* Oxford: Pergamon Press.

Spearritt, D (Ed) (1982) *The Improvement of Measurement in Education and Psychology.* Hawthorn, Victoria: Australian Council for Educational Research.

Stenner, A J, Smith, M & Burdick, D S (1983) Toward a theory of construct definition. *Journal of Educational Measurement*, 20,4,305-316.

Wainer, H & Messick, S (Eds) (1983) *Principals (sic) of Modern Psychological Measurement.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Wright, B D (1980) Foreword and Afterword in Rasch, G *Probabilistic Models for some Intelligence and Attainment Tests.* (Expanded edition) Chicago: University of Chicago Press.

APPENDIX A
RASCH STATISTICS: METHODS OF CALCULATION

Adapted from Wright & Stone (1979), Wright, Mead & Bell (1980) and Wright & Masters (1982)

A.1 Item Difficulty and Person Ability Estimates (UCON)

(i) After removal of persons scoring 0 or full marks, and items answered all correctly or all incorrectly, initial ability estimates are calculated from

$$b_r^{(0)} = \ln[L/(L - r)]$$

where $b_r^{(0)}$ = initial ability estimate for persons with raw score r
r = raw score
L = number of items.

(ii) Initial difficulty estimates are calculated from

$$d_i^{(0)} = \ln[(N-s_i)/s_i]$$

where $d_i^{(0)}$ = initial difficulty estimate for item i
s_i = number-correct score for item i
N = number of persons.

(iii) Item difficulties are centred by subtracting their mean

$$d_. = \sum_{i=1}^{L} d_i/L$$

from each difficulty estimate.

(iv) Set of item difficulties is then revised using the iterative formula

$$d_i^{(j+1)} = d_i^{(j)} - \frac{s_i - \sum_{r=1}^{L-1} n_r p_{ri}^{(j)}}{\sum_{r=1}^{L-1} n_r p_{ri}^{(j)}(1 - p_{ri}^{(j)})}$$

where n_r = number of persons in raw score group r
p_{ri} = probability that a person with raw score r will succeed on item i
$p_{ri}^{(j)} = \exp(b_r - d_i^{(j)})/[1 + \exp(b_r - d_i^{(j)})]$.

The superscript $^{(j)}$ denotes the iteration number. Iteration continues until successive estimates differ by less than 0.01, i.e.

$$|d_i^{(j+1)} - d_i^{(j)}| < 0.01$$

for each item i.

(v) Using the revised values d_i, the set of ability estimates is revised using the iterative formula

$$b_r^{(m+1)} = b_r^{(m)} + \frac{r - \sum_{i=1}^{L} p_{ri}^{(m)}}{\sum_{i=1}^{L} p_{ri}^{(m)}(1 - p_{ri}^{(m)})}$$

until convergence is reached at

$$|b_r^{(m+1)} - b_r^{(m)}| < 0.01$$

for each raw score group r, where

$$p_{ri}^{(m)} = \exp(b_r^{(m)} - d_i)/[1 + \exp(b_r^{(m)} - d_i)] .$$

A.2 Standard Errors of Difficulty and Ability Estimates (UCON)

Using the estimated difficulties d_i and abilities b_r, standard errors (SE) are calculated from

$$SE(d_i) = \left| \sum_{r=1}^{L-1} n_r p_{ri}(1 - p_{ri}) \right|^{-\frac{1}{2}}$$

for each item i, and

$$SE(b_r) = \left| \sum_{i=1}^{L} p_{ri}(1 - p_{ri}) \right|^{-\frac{1}{2}}$$

for each raw score group r, where

$$p_{ri} = \exp(b_r - d_i)/[1 + \exp(b_r - d_i)]$$

A.3 Information-Weighted Total Fit t-Statistics for Persons and Items

For each person-item response, model expectation is calculated from

$$p_{vi} = \exp(b_v - d_i)/[1 + \exp(b_v - d_i)]$$

where b_v = estimated ability of person v
d_i = estimated difficulty of item i.

Residuals for each person-item response are calculated from

$$x_{vi} - p_{vi}$$

where x_{vi} = observed response (1 = correct, 0 = incorrect) of person v on item i.

These residuals are then squared and summed over persons for item fit:

$$\sum_{v=1}^{N} (x_{vi} - p_{vi})^2$$

and over items for person fit:

$$\sum_{i=1}^{L} (x_{vi} - p_{vi})^2 .$$

Mean square statistics are then formed for items from

$$w_i = \sum_{v=1}^{N} (x_{vi} - p_{vi})^2 / [\sum_{v=1}^{N} p_{vi}(1-p_{vi})]$$

and for persons from

$$w_v = \sum_{i=1}^{L} (x_{vi} - p_{vi})^2 / [\sum_{i=1}^{L} p_{vi}(1-p_{vi})]$$

with corresponding variances

$$s_i^2 = (\sum_{v=1}^{N} p_{vi}(1 - p_{vi}) - 4 \sum_{v=1}^{N} [p_{vi}(1 - p_{vi})]^2) / [\sum_{v=1}^{N} p_{vi}(1 - p_{vi})]^2$$

and

$$s_v^2 = (\sum_{i=1}^{L} p_{vi}(1 - p_{vi}) - 4 \sum_{i=1}^{L} [p_{vi}(1 - p_{vi})]^2) / [\sum_{i=1}^{L} p_{vi}(1 - p_{vi})]^2 .$$

Finally, the total fit t-statistics are obtained from

$$t_i = 3(w_i^{1/3} - 1)/s_i + s_i/3$$

for each item i , and

$$t_v = 3(w_v^{1/3} - 1)/s_v + s_v/3$$

for each person v .

A.4 Between-Group Fit t-Statistics for Items

The sample is first divided into a number (in this study, 6) of roughly equal sized subgroups on the basis of score level. For each subgroup g on each item i, the difference between observed number of correct answers and expected number for that subgroup is calculated from

$$s_{gi} - \sum_{r \in g} n_r p_{ri}$$

where s_{gi} = observed number of correct answers on item i in subgroup g
r = raw score
$r \in g$ = for each raw score in subgroup g
n_r = number of persons with raw score r
p_{ri} = model probability that a person with raw score r will succeed on item i .

Standardized residuals are then calculated from

$$z_{gi} = [s_{gi} - \sum_{r \in g} n_r p_{ri}] / [\sum_{r \in g} n_r p_{ri} (1-p_{ri})]^{\frac{1}{2}}$$

and a mean square between the M subgroups from

$$w_{Bi} = \sum_{g=1}^{M} z_{gi}^2 [L/((M-1)(L-1))] .$$

Finally, the between-group fit t-statistic for each item i is calculated from

$$t_{Bi} = a w_{Bi}^{1/3} - a + 1/a$$

where $a = [4.5(M-1)]^{\frac{1}{2}}$.

A.5 Rasch-Based Discrimination Index

The mean ability of the persons in each of the M (here, 6) score groups is calculated:

$$b_{.g} = \sum_{r \in g} n_r b_r / N_g$$

where n_r = no. of persons with raw score r in group g
b_r = ability corresponding to raw score r
N_g = no. of persons in group g.

These means are centred about zero by subtracting from each the mean ability of the whole sample ($b_.$):

$$b_. = \sum_{v=1}^{N} b_v / N$$

to give M values ($b_{.g} - b_.$). 'Group residuals' are calculated in the form

$$(s_{gi} - \sum_{r \in g} n_r p_{ri})$$

where s_{gi} is the observed score for group g on item i.

These terms are then scaled by the centred abilities and summed over the groups:

$$X = \sum_{g=1}^{M} (b_{.g} - b_.)(s_{gi} - \sum_{r \in g} n_r p_{ri}) .$$

The variance for each group g is calculated:

$$\sum_{r \in g} n_r p_{ri}(1-p_{ri}) .$$

These values are then scaled by the square of centred abilities and summed over the groups to give

$$Y = \sum_{g=1}^{M} (b_{.g} - b)^2 [\sum_{r \in g} n_r p_{ri}(1-p_{ri})] .$$

From the above, the Rasch discrimination index for a given item i (a_i) is calculated from the formula

$$a_i = X/Y + 1 .$$

A.6 Person Separability Index and No. of Person Strata

The variance of the person ability estimates is first calculated. A mean square measurement error is calculated from the standard errors of estimated ability:

$$MSE = \sum_{v=1}^{N} SE(b_v)^2 / N$$

The ability variance is adjusted for the measurement error to give an adjusted sample variance:

$$SA^2 = SD^2 - MSE$$

A separation index is calculated from the adjusted sample standard deviation divided by the root mean square of measurement error:

$$SI = SA/\sqrt{MSE}$$

The person separability index RI (or test reliability of person separation), which shows the proportion of observed sample variance not due to measurement error, is calculated from

$$RI = SA^2/(SA^2 + MSE)$$

The number of person strata, i.e. the number of distinct levels of person ability separated by a distance of 3 measurement errors, is calculated as follows:

$$\text{No. of strata} = (4SI + 1)/3$$

APPENDIX B
CLOZE-TYPE TEST: TEST PAPER AND MARKING SHEET

B.1 Cloze-Type Test Paper

A There is a rambutan tree in Lalita's garden. It (1) a big tree. Swami likes (2) climb the tree. One day (3) climbed very high. He looked (4) the garden wall and (5) all the fields around. "I (6) see some buffalo in the river," Swami called (7) Lalita. "What else can (8) see?" asked Lalita. "I can (9) two dogs on the road." " (10) are they doing?" asked Lalita. " (11) are fighting".

B Jenny Lim and her brother Peter went for a walk. As (12) passed the big house on (13) hill a dog ran out. It (14) a small brown dog (15) short legs. It was barking (16) . Jenny was frightened. But Peter (17) "It is only a small (18) . Don't be frightened." He picked (19) a stone and threw it (20) the dog. It ran towards Jenny (21) tried to bite her. Peter saw (22) big stick and picked (23) up. The dog quickly ran away up (24) hill.

C Old Mrs. Chong lived in a small house at (25) end of a village. Her (26) was dead. Her children were (27) up and lived a (28) way away. So she was (29) . No one in the village liked (30) . No one came to visit her. (31) day Mrs. Seng came to her (32) and said "Mrs. Chong, may (33) come in? I've brought you (34) fruit. I said to myself, 'I must (35) Mrs. Chong some of my (36) . She hasn't got any in (37) garden'."

D Yesterday was Ali's birthday. His mother and father took (38) to the zoo. They went (39) the morning by bus. They took (40) food with them. Ali liked all (41) animals. They went round and (42) the zoo looking at the animals. (43) about 11.30 Ali felt (44) hungry. Then he sat down (45) his mother and father under (46) tree. It was very cool (47) the tree. They ate (48) big lunch. After lunch they (49) to sleep. Later, a noise (50) them up. Ali felt very happy (51) pleased with his visit (52) the zoo.

E Tony followed Mitch through the dark hole. "We're (53) a tunnel", Tony said. "Someone (54) cut it in the rock. Where (55) we now?" "This is part (56) an old mine," said Mitch. "Maybe it isn't very safe. So (57) must walk carefully here. Stay (58) me," "Yes, but look at (59) rock!" Tony dropped his rope (60) climbed over some big rocks. " (61) on, Mitch. Bring your torch (62) here." Mitch shone his (63) on the roof of the (64) .

F Mr. Davey was a very old man and he (65) very curious. His eyes (66) still good and his ears were (67) too. He could see a (68) of things and he could (69) a lot of things. He (70) sitting in the porch of (71) daughter's house, and he was talking to (72) . She was sweeping the floor inside (73) house. "Look, old Mrs. Benson is (74) into that shop again. It's the fifth (75) today that the old lady (76) gone in there."

G When Peter was young, he fell ill and lay (77) bed unconscious. Doctors, of course, (78) their best for him and (79) to make him better; but (80) remained unconscious for a long (81) . Then he suddenly began to (82) clearly. He described the cause (83) his illness and explained all (84) things that must be done (85) make him better. The doctors (86) as the boy said, and (87) soon began to get better.

H Then suddenly I had a wonderful idea. Every morning (88) half-past six the milkman (89) my milk. He was a short man and we were the (90) size. He had a short (91) moustache and wore a white (92) and coat. My idea was (93) borrow his clothes and the (94) of milk. Then I could (95) from the building as the (96) . No one watching would know it (97) me.

I Tun Perak and his companies watched the darkness, (98) until they were sure (99) the Siamese were fast (100) . Silently Tun crept (101) on the man on guard, as (102) stood looking in the direction (103) the river. With a (104) thrust he drove his sharp kris (105) the guard's heart from the (106) while his left hand covered (107) man's nose and mouth. The (108) fell to the ground without a sound.

J Gwen put her hand on his arm and looked (109) his face. "What is worrying you, David?" "Something (110) silly. It's difficult to explain. I (111) a fool. "But what's (112) ?" she asked with more force. " (113) feel like a man trying (114) remember something. Have I forgotten (115) about the reactor? Could there (116) any danger there?" "An explosion?" (117) asked. "No. An explosion couldn't (118) . The reactor isn't even like (119) explosion. It's like a slow fire."

K "I don't see the point of it," said Micky. (120) were both laughing. They were (121) sure why they were laughing. (122) it was just for fun (123) because they were young. Harold (124) on rowing. The sun was (125) . The fields on the opposite bank (126) bright and they could (127) the farmhouse in which they were (128) . Its windows reflected the evening (129) .

L The Air Hostess went away and came back with a (130) of whisky. She seemed (131) . She had blue eyes. He wished he could be as calm (132) she appeared to be. The plane's (133) grew quieter. For a moment they (134) to have stopped completely. The (135) dropped like a stone, and (136) dived into the grey clouds. He (137) see nothing except a (138) white mist outside the windows. (139) in the plane was talking to each (140) . The plane seemed to fall (141) and down.

B.2 Marking Sheet for Cloze-Type Test

A. 1. is
2. to
3. he/Swami
4. over
5. at/into/across/saw/observed/viewed
6. can
7. to
8. you
9. see
10. what
11. they

B. 12. they
13. the
14. was
15. with
16. loudly/angrily/fiercely/noisily/etc.
17. said
18. dog/puppy/one
19. up
20. at
21. and
22. a
23. it
24. the

C. 25. the,one
26. husband
27. grown
28. long
29. alone/lonely
30. her
31. one
32. house/neighbour/quietly/door
33. I
34. some
35. give/take
36. fruit(s)/apples etc.
37. her

D. 38. him
39. in
40. some
41. the
42. round
43. at
44. very/quite/extremely
45. with/beside/between/near
46. a
47. under
48. a
49. went
50. woke
51. and
52. to

E. 53. in(side)
54. has
55. are
56. of
57. we
58. with/behind/near/beside/by
59. that/the
60. and
61. come
62. over/up/down
63. torch/torchlight/light
64. tunnel

F. 65. was
66. were
67. good
68. lot
69. hear
70. was
71. his
72. her
73. the/her
74. going
75. time
76. has

G. 77. in
78. did
79. tried/attempted
80. he/Peter
81. time/period
82. think/talk/speak
83. of
84. the
85. to
86. did
87. he/Peter

H. 88. at
89. delivered, brought
90. same
91. (colour)/bushy/curly/untidy/thick etc.
92. cap/hat
93. to
94. crate/bottle(s)
95. escape/emerge/walk
96. milkman
97. was

I. 98. right/waiting
99. that
100. asleep
101. up
102. he
103. of
104. sharp/sudden/quick/strong/deadly/silent/mighty/fast etc.
105. into/through
106. right/back/side
107. the
108. man/guard/victim

J. 109. into
110. very/rather/quite/really
111. feel
112. up/wrong
113. I
114. to
115. something/anything
116. be
117. she/Gwen
118. happen/occur
119. an

K. 120. they
121. not
122. perhaps/maybe/either
123. or/and
124. went/carried/kept
125. low/setting/shining/sinking
126. were/looked
127. see
128. staying/living
129. sun(light)/sky/sunset/light

L. 130. glass
131. calm/cool/friendly etc.
132. as
133. engine(s)
134. seemed
135. (aero)plane
136. then/suddenly
137. could
138. thick/dense
139. everyone (body)
140. other
141. down

APPENDIX C
TRADITIONAL STATISTICS FOR CLOZE-TYPE TEST (MALAYSIAN DATA)

C.1 Cloze-Type Test (Malaysian Group): Raw Score Distribution & Frequency Counts, K-R20 & SEM

RAW SCORE	FREQ.	CUM. FREQ.	NO. OF PERSONS 0 | 10 | 20
0	3	3	***
1	1	4	*
2	0	4	
3	0	4	
4	0	4	
5	1	5	*
6	1	6	*
7	1	7	*
8	2	9	**
9	0	9	
10	1	10	*
11	2	12	**
12	1	13	*
13	3	16	***
14	2	18	**
15	2	20	**
16	1	21	*
17	4	25	****
18	1	26	*
19	2	28	**
20	0	28	
21	2	30	**
22	2	32	**
23	3	35	***
24	6	41	******
25	1	42	*
26	4	46	****
27	1	47	*
28	2	49	**
29	2	51	**
30	1	52	*
31	2	54	**
32	4	58	****
33	1	59	*
34	3	62	***
35	3	65	***
36	2	67	**
37	1	68	*
38	2	70	**
39	6	76	******
40	2	78	**
41	3	81	***
42	1	82	*
43	2	84	**
44	1	85	*
45	2	87	**
46	5	92	*****
47	3	95	***
48	4	99	****
49	3	102	***
50	4	106	****

RAW SCORE	FREQ.	CUM. FREQ.	NO. OF PERSONS 0 | 10 | 20 |
51	1	107	*
52	4	111	****
53	2	113	**
54	3	116	***
55	6	122	******
56	1	123	*
57	1	124	*
58	4	128	****
59	2	130	**
60	3	133	***
61	7	140	*******
62	2	142	**
63	6	148	******
64	1	149	*
65	4	153	****
66	8	161	********
67	6	167	******
68	6	173	******
69	4	177	****
70	4	181	****
71	7	188	*******
72	5	193	*****
73	2	195	**
74	6	201	******
75	5	206	*****
76	5	211	*****
77	3	214	***
78	11	225	**********
79	5	230	*****
80	6	236	******
81	4	240	****
82	8	248	********
83	7	255	*******
84	9	264	*********
85	9	273	*********
86	6	279	******
87	5	284	*****
88	3	287	***
89	2	289	**
90	2	291	**
91	5	296	*****
92	7	303	*******
93	7	310	*******
94	4	314	****
95	3	317	***
96	4	321	****
97	8	329	********
98	8	337	********
99	9	346	*********
100	5	351	*****

RAW SCORE	FREQ.	CUM. FREQ.	NO. OF PERSONS 0 10 20 | | |
101	3	354	***
102	7	361	*******
103	8	369	********
104	14	383	**************
105	9	392	*********
106	12	404	************
107	4	408	****
108	7	415	*******
109	5	420	*****
110	8	428	********
111	11	439	***********
112	7	446	*******
113	4	450	****
114	7	457	*******
115	10	467	**********
116	6	473	******
117	8	481	********
118	16	497	*****************
119	6	503	******
120	11	514	***********
121	8	522	********
122	8	530	********
123	15	545	***************
124	7	552	*******
125	11	563	***********
126	8	571	********
127	9	580	*********
128	7	587	*******
129	3	590	***
130	6	596	******
131	6	602	******
132	1	603	*
133	5	608	*****
134	2	610	**
135	0	610	|
136	1	611	*
137	0	611	|
138	0	611	|
139	0	611	|
140	0	611	|
141	0	611	|

Mean raw score = 86.20
Standard deviation = 33.36
Raw score range for group: 0 to 136

K-R20 = 0.98
SEM = 4.22

C.2 Cloze-Type Test (Malaysian Group): Traditional Item Statistics

ITEM LABEL	FACILITY VALUE	DISCRIM INDEX	UNBIASED PT.BISERIAL
A 1	0.94	0.16	0.29
A 2	0.94	0.24	0.50
A 3	0.90	0.32	0.49
A 4	0.36	0.76	0.58
A 5	0.46	0.82	0.62
A 6	0.77	0.45	0.44
A 7	0.56	0.60	0.39
A 8	0.90	0.28	0.44
A 9	0.92	0.23	0.44
A 10	0.94	0.21	0.46
A 11	0.90	0.29	0.45
B 12	0.84	0.53	0.66
B 13	0.77	0.09	0.10
B 14	0.66	0.69	0.55
B 15	0.75	0.80	0.77
B 16	0.51	0.70	0.54
B 17	0.50	0.62	0.45
B 18	0.89	0.30	0.44
B 19	0.84	0.47	0.57
B 20	0.52	0.85	0.68
B 21	0.86	0.48	0.64
B 22	0.87	0.36	0.50
B 23	0.87	0.46	0.64
B 24	0.77	0.52	0.52
C 25	0.95	0.12	0.33
C 26	0.84	0.51	0.64
C 27	0.51	0.90	0.69
C 28	0.24	0.57	0.47
C 29	0.70	0.82	0.72
C 30	0.84	0.49	0.61
C 31	0.91	0.21	0.35
C 32	0.84	0.36	0.44
C 33	0.94	0.19	0.45
C 34	0.81	0.16	0.27
C 35	0.65	0.61	0.53
C 36	0.66	0.47	0.41
C 37	0.68	0.76	0.66
D 38	0.87	0.42	0.58
D 39	0.74	0.61	0.59
D 40	0.80	0.35	0.40
D 41	0.89	0.33	0.47
D 42	0.66	0.65	0.56
D 43	0.75	0.59	0.58
D 44	0.78	0.30	0.33
D 45	0.80	0.55	0.59
D 46	0.65	0.86	0.73
D 47	0.85	0.36	0.47
D 48	0.72	0.52	0.46
D 49	0.65	0.73	0.59
D 50	0.53	0.93	0.72

ITEM LABEL	FACILITY VALUE	DISCRIM INDEX	UNBIASED PT.BISERIAL
D 51	0.85	0.51	0.61
D 52	0.87	0.36	0.47
E 53	0.70	0.68	0.63
E 54	0.39	0.55	0.40
E 55	0.78	0.47	0.51
E 56	0.82	0.59	0.68
E 57	0.83	0.50	0.60
E 58	0.82	0.53	0.61
E 59	0.30	0.68	0.52
E 60	0.84	0.42	0.52
E 61	0.59	0.75	0.61
E 62	0.45	0.84	0.63
E 63	0.65	0.73	0.63
E 64	0.40	0.67	0.55
F 65	0.73	0.39	0.33
F 66	0.50	0.81	0.61
F 67	0.72	0.35	0.37
F 68	0.54	0.75	0.57
F 69	0.54	0.80	0.64
F 70	0.68	0.62	0.54
F 71	0.79	0.53	0.60
F 72	0.74	0.65	0.59
F 73	0.82	0.25	0.29
F 74	0.61	0.79	0.64
F 75	0.65	0.87	0.72
F 76	0.45	0.84	0.63
G 77	0.47	0.81	0.63
G 78	0.36	0.87	0.63
G 79	0.44	0.85	0.62
G 80	0.80	0.53	0.57
G 81	0.82	0.39	0.44
G 82	0.30	0.48	0.41
G 83	0.82	0.58	0.66
G 84	0.87	0.38	0.55
G 85	0.84	0.49	0.61
G 86	0.37	0.93	0.67
G 87	0.67	0.73	0.64
H 88	0.79	0.49	0.56
H 89	0.17	0.35	0.33
H 90	0.65	0.68	0.58
H 91	0.44	0.88	0.66
H 92	0.20	0.40	0.35
H 93	0.82	0.56	0.65
H 94	0.40	0.55	0.43
H 95	0.22	0.45	0.38
H 96	0.66	0.82	0.70
H 97	0.60	0.72	0.55
I 98	0.08	0.18	0.23
I 99	0.73	0.64	0.61
I100	0.47	0.92	0.72

ITEM LABEL	FACILITY VALUE	DISCRIM INDEX	UNBIASED PT.BISERIAL
I101	0.18	0.44	0.36
I102	0.72	0.76	0.70
I103	0.63	0.82	0.69
I104	0.31	0.65	0.52
I105	0.48	0.92	0.68
I106	0.53	0.79	0.64
I107	0.75	0.75	0.71
I108	0.71	0.75	0.68
J109	0.09	0.23	0.30
J110	0.30	0.71	0.55
J111	0.05	0.13	0.19
J112	0.27	0.52	0.42
J113	0.72	0.59	0.55
J114	0.86	0.45	0.62
J115	0.28	0.53	0.43
J116	0.64	0.92	0.78
J117	0.55	0.67	0.53
J118	0.35	0.75	0.55
J119	0.60	0.78	0.62
K120	0.71	0.61	0.54
K121	0.51	0.70	0.53
K122	0.22	0.63	0.51
K123	0.36	0.56	0.43
K124	0.31	0.77	0.61
K125	0.51	0.72	0.56
K126	0.39	0.75	0.57
K127	0.69	0.82	0.72
K128	0.30	0.68	0.55
K129	0.54	0.78	0.60
L130	0.76	0.61	0.61
L131	0.31	0.82	0.61
L132	0.67	0.88	0.76
L133	0.11	0.37	0.39
L134	0.26	0.75	0.58
L135	0.62	0.86	0.71
L136	0.18	0.28	0.27
L137	0.59	0.93	0.75
L138	0.18	0.46	0.43
L139	0.12	0.31	0.33
L140	0.73	0.61	0.59
L141	0.54	0.58	0.46

No. of persons = 611

C.3 Grouped Item Statistics (Malaysian Data)

Table 1 Cloze-Type Test (Malaysian Data): Items Grouped by Facility Value

Facility Value Interval	No.of Items	Item Names
0.90-1.00	10	A1 A2 A3 A8 A9 A10 A11 C25 C31 C33
0.80-0.89	29	B12 B18 B19 B21 B22 B23 C26 C30 C32 C34 D38 D40 D41 D45 D47 D51 D52 E56 E57 E58 E60 F73 G80 G81 G83 G84 G85 H93 J114
0.70-0.79	24	A6 B13 B15 B24 C29 D39 D43 D44 D48 E53 E55 F65 F67 F71 F72 H88 I99 I102 I107 I108 J113 K120 L130 L140
0.60-0.69	21	B14 C35 C36 C37 D42 D46 D49 E63 F70 F74 F75 G87 H90 H96 H97 I103 J116 J119 K127 L132 L135
0.50-0.59	17	A7 B16 B17 B20 C27 D50 E61 F66 F68 F69 I106 J117 K121 K125 K129 L137 L141
0.40-0.49	10	A5 E62 E64 F76 G77 G79 H91 H94 I100 I105
0.30-0.39	14	A4 E54 E59 G78 G82 G86 I104 J110 J118 K123 K124 K126 K128 K131
0.20-0.29	7	C28 H92 H95 J112 J115 K122 L134
0.10-0.19	6	H89 I101 L133 L136 L138 L139
0.00-0.09	3	I98 J109 J111

Facility value range = 0.05 (Item J111) to 0.95 (Item C25)
Mean = 0.61
SD = 0.23

Table 2 Cloze-Type Test (Malaysian Data): Items Grouped by Discrimination Index

Discrim. Index Interval	No.of Items	Item Names
0.90-1.00	7	C27 D50 G86 I100 I105 J116 L137
0.80-0.89	20	A5 B15 B20 C29 D46 E62 F66 F69 F75 F76 G77 G78 G79 H91 H96 I103 K127 L131 L132 L135
0.70-0.79	23	A4 B16 C37 D49 E61 E63 F68 F74 G87 H97 I102 I106 I107 I108 J110 J118 J119 K121 K124 K125 K126 K129 L134
0.60-0.69	20	A7 B14 B17 C35 D39 D42 E53 E59 E64 F70 F72 H90 I99 I104 J117 K120 K122 K128 L130 L140
0.50-0.59	22	B12 B24 C26 C28 D43 D45 D48 D51 E54 E56 E57 E58 F71 G80 G83 H93 H94 J112 J113 J115 K123 L141
0.40-0.49	17	A6 B19 B21 B23 C30 C36 D38 E55 E60 G82 G85 H88 H92 H95 I101 J114 L138
0.30-0.39	16	A3 B18 B22 C32 D40 D41 D44 D47 D52 F65 F67 G81 G84 H89 L133 L139
0.20-0.29	9	A2 A8 A9 A10 A11 C31 F73 J109 L136
0.10-0.19	6	A1 C25 C33 C34 I98 J111
0.00-0.09	1	B13

Discrimination index range = 0.09 (Item B13) to 0.93 (Item D50)
Mean = 0.58
SD = 0.21

Table 3 Cloze-Type Test (Malaysian Data): Items Grouped by Unbiased Point Biserial

Point Biserial Interval	No.of Items	Item Names
0.90-1.00	0	
0.80-0.89	0	
0.70-0.79	14	B15 C29 D46 D50 F75 H96 I100 I102 I107 J116 K127 L132 L135 L137
0.60-0.69	42	A5 B12 B20 B21 B23 C26 C27 C30 C37 D51 E53 E56 E57 E58 E61 E62 E63 F66 F69 F71 F74 F76 G77 G78 G79 G83 G85 G86 G87 H91 H93 I99 I103 I105 I106 I108 J114 J119 K124 K129 L130 L131
0.50-0.59	39	A2 A4 B14 B16 B19 B22 B24 C35 D38 D39 D42 D43 D45 D49 E55 E59 E60 E64 F68 F70 F72 G80 G84 H88 H90 H97 I104 J110 J113 J117 J118 K120 K121 K122 K125 K126 K128 L134 L140
0.40-0.49	26	A3 A6 A8 A9 A10 A11 B17 B18 C28 C32 C33 C36 D40 D41 D47 D48 D52 E54 G81 G82 H94 J112 J115 K123 L138 L141
0.30-0.39	13	A7 C25 C31 D44 F65 F67 H89 H92 H95 I101 J109 L133 L139
0.20-0.29	5	A1 C34 F73 I98 L136
0.10-0.19	2	B13 J111
0.00-0.09	0	

Point biserial range = 0.10 (Item B13) to 0.78 (Item J116)
Mean = 0.54
SD = 0.13

C.4 Cloze-Type Test (Whole Malaysian Group): Item Z-Scores & Z-Scale Values

ITEM NAME	ITEM Z-SCORE	ITEM Z-SCALE VALUE
A 1	1.41	-1.55
A 2	1.40	-1.52
A 3	1.24	-1.28
A 4	-1.08	0.35
A 5	-0.67	0.11
A 6	0.68	-0.73
A 7	-0.22	-0.15
A 8	1.25	-1.29
A 9	1.35	-1.44
A 10	1.41	-1.55
A 11	1.25	-1.29
B 12	0.99	-1.00
B 13	0.68	-0.73
B 14	0.20	-0.41
B 15	0.61	-0.68
B 16	-0.44	-0.02
B 17	-0.46	-0.01
B 18	1.18	-1.20
B 19	0.97	-0.98
B 20	-0.40	-0.05
B 21	1.08	-1.09
B 22	1.12	-1.14
B 23	1.10	-1.11
B 24	0.69	-0.74
C 25	1.48	-1.69
C 26	0.99	-1.00
C 27	-0.43	-0.03
C 28	-1.62	0.72
C 29	0.37	-0.52
C 30	0.99	-1.00
C 31	1.29	-1.34
C 32	1.01	-1.02
C 33	1.43	-1.58
C 34	0.84	-0.87
C 35	0.18	-0.39
C 36	0.20	-0.41
C 37	0.29	-0.47
D 38	1.11	-1.12
D 39	0.56	-0.64
D 40	0.83	-0.86
D 41	1.21	-1.24
D 42	0.21	-0.41
D 43	0.62	-0.69
D 44	0.72	-0.77
D 45	0.80	-0.83
D 46	0.18	-0.39
D 47	1.03	-1.04
D 48	0.45	-0.57
D 49	0.16	-0.38
D 50	-0.35	-0.08

ITEM NAME	ITEM Z-SCORE	ITEM Z-SCALE VALUE
D 51	1.03	-1.03
D 52	1.10	-1.11
E 53	0.37	-0.51
E 54	-0.94	0.27
E 55	0.73	-0.78
E 56	0.88	-0.90
E 57	0.94	-0.95
E 58	0.90	-0.92
E 59	-1.34	0.52
E 60	0.99	-1.00
E 61	-0.07	-0.24
E 62	-0.67	0.11
E 63	0.16	-0.38
E 64	-0.89	0.24
F 65	0.50	-0.61
F 66	-0.46	-0.01
F 67	0.45	-0.57
F 68	-0.30	-0.11
F 69	-0.32	-0.09
F 70	0.31	-0.48
F 71	0.78	-0.81
F 72	0.55	-0.64
F 73	0.89	-0.90
F 74	-0.02	-0.27
F 75	0.16	-0.38
F 76	-0.70	0.13
G 77	-0.61	0.07
G 78	-1.08	0.35
G 79	-0.74	0.15
G 80	0.82	-0.85
G 81	0.88	-0.90
G 82	-1.35	0.53
G 83	0.89	-0.90
G 84	1.10	-1.11
G 85	0.99	-1.00
G 86	-1.03	0.33
G 87	0.25	-0.44
H 88	0.78	-0.81
H 89	-1.90	0.95
H 90	0.19	-0.40
H 91	-0.72	0.14
H 92	-1.75	0.82
H 93	0.88	-0.90
H 94	-0.91	0.25
H 95	-1.70	0.78
H 96	0.23	-0.42
H 97	-0.04	-0.25
I 98	-2.30	1.41
I 99	0.51	-0.61
I100	-0.61	0.07

ITEM NAME	ITEM Z-SCORE	ITEM Z-SCALE VALUE
I101	-1.86	0.92
I102	0.48	-0.59
I103	0.09	-0.34
I104	-1.28	0.49
I105	-0.58	0.06
I106	-0.36	-0.07
I107	0.58	-0.67
I108	0.43	-0.55
J109	-2.27	1.37
J110	-1.33	0.51
J111	-2.42	1.64
J112	-1.45	0.60
J113	0.45	-0.57
J114	1.07	-1.08
J115	-1.41	0.57
J116	0.10	-0.35
J117	-0.24	-0.14
J118	-1.13	0.39
J119	-0.04	-0.25
K120	0.40	-0.54
K121	-0.43	-0.03
K122	-1.70	0.78
K123	-1.10	0.37
K124	-1.29	0.49
K125	-0.45	-0.02
K126	-0.95	0.28
K127	0.33	-0.49
K128	-1.33	0.44
K129	-0.33	-0.09
L130	0.65	-0.72
L131	-1.30	0.49
L132	0.27	-0.45
L133	-2.14	1.20
L134	-1.50	0.63
L135	0.04	-0.31
L136	-1.85	0.90
L137	-0.08	-0.23
L138	-1.85	0.90
L139	-2.14	1.20
L140	0.51	-0.61
L141	-0.31	-0.10

No. of persons = 611

C.5 Cloze-Type Test (High Scorers, Malaysia): Facility Values, Item Z-Scores & Item Z-Scale Values

ITEM NAME	FACILITY VALUE	ITEM Z-SCORE	ITEM Z-SCALE VALUE
A 1	0.99	0.72	-2.58
A 2	1.00	0.74	-4.00
A 3	0.99	0.72	-2.58
A 4	0.76	-0.47	-0.71
A 5	0.88	0.14	-1.18
A 6	0.95	0.49	-1.64
A 7	0.91	0.31	-1.37
A 8	0.99	0.69	-2.33
A 9	0.99	0.69	-2.33
A 10	1.00	0.74	-4.00
A 11	1.00	0.74	-4.00
B 12	1.00	0.74	-4.00
B 13	0.84	-0.06	-0.99
B 14	0.96	0.57	-1.10
B 15	1.00	0.74	-4.00
B 16	0.78	-0.37	-0.77
B 17	0.75	-0.49	-0.69
B 18	0.99	0.69	-2.33
B 19	1.00	0.74	-4.00
B 20	0.97	0.59	-1.88
B 21	1.00	0.74	-4.00
B 22	0.99	0.72	-2.58
B 23	1.00	0.74	-4.00
B 24	0.97	0.59	-1.88
C 25	0.99	0.69	-2.33
C 26	1.00	0.74	-4.00
C 27	0.92	0.36	-1.44
C 28	0.54	-1.56	-0.11
C 29	0.98	0.64	-2.05
C 30	0.98	0.67	-2.17
C 31	0.98	0.67	-2.17
C 32	0.97	0.62	-1.96
C 33	1.00	0.74	-4.00
C 34	0.82	-0.14	-0.93
C 35	0.89	0.19	-1.23
C 36	0.78	-0.37	-0.77
C 37	0.97	0.59	-1.88
D 38	0.99	0.69	-2.33
D 39	0.95	0.49	-1.64
D 40	0.92	0.34	-1.41
D 41	0.99	0.72	-2.58
D 42	0.89	0.21	-1.25
D 43	0.96	0.57	-1.81
D 44	0.88	0.14	-1.18
D 45	0.98	0.64	-2.05
D 46	0.97	0.62	-1.96
D 47	0.97	0.62	-1.96
D 48	0.93	0.42	-1.51
D 49	0.96	0.54	-1.75
D 50	0.97	0.59	-1.88

ITEM NAME	FACILITY VALUE	ITEM Z-SCORE	ITEM Z-SCALE VALUE
D 51	0.99	0.72	-2.58
D 52	0.99	0.69	-2.33
E 53	0.94	0.47	-1.60
E 54	0.65	-1.00	-0.40
E 55	0.99	0.69	-2.33
E 56	0.99	0.69	-2.33
E 57	0.98	0.67	-2.17
E 58	0.97	0.62	-1.96
E 59	0.65	-1.02	-0.39
E 60	0.98	0.64	-2.05
E 61	0.91	0.31	-1.37
E 62	0.87	0.09	-1.13
E 63	0.96	0.54	-1.75
E 64	0.72	-0.67	-0.58
F 65	0.91	0.29	-1.34
F 66	0.87	0.09	-1.13
F 67	0.77	-0.39	-0.76
F 68	0.89	0.19	-1.23
F 69	0.87	0.09	-1.13
F 70	0.95	0.52	-1.70
F 71	0.97	0.62	-1.96
F 72	0.96	0.57	-1.81
F 73	0.91	0.29	-1.34
F 74	0.96	0.54	-1.75
F 75	0.99	0.69	-2.33
F 76	0.88	0.14	-1.18
G 77	0.90	0.24	-1.28
G 78	0.84	-0.06	-0.99
G 79	0.86	0.06	-1.10
G 80	0.99	0.69	-2.33
G 81	0.98	0.64	-2.05
G 82	0.50	-1.76	-0.01
G 83	0.98	0.67	-2.17
G 84	0.99	0.69	-2.33
G 85	1.00	0.74	-4.00
G 86	0.88	0.11	-1.15
G 87	0.93	0.42	-1.51
H 88	0.98	0.64	-2.05
H 89	0.34	-2.59	0.41
H 90	0.93	0.39	-1.48
H 91	0.88	0.14	-1.18
H 92	0.40	-2.26	0.24
H 93	0.99	0.72	-2.58
H 94	0.64	-1.08	-0.36
H 95	0.43	-2.14	0.18
H 96	0.96	0.57	-1.81
H 97	0.86	0.06	-1.10
I 98	0.18	-3.40	0.92
I 99	0.91	0.31	-1.37
I100	0.90	0.26	-1.31

ITEM NAME	FACILITY VALUE	ITEM Z-SCORE	ITEM Z-SCALE VALUE
I101	0.42	-2.16	0.19
I102	0.98	0.67	-2.17
I103	0.95	0.49	-1.64
I104	0.64	-1.08	-0.36
I105	0.92	0.36	-1.44
I106	0.87	0.09	-1.13
I107	0.99	0.72	-2.58
I108	0.97	0.59	-1.88
J109	0.21	-3.25	0.81
J110	0.68	-0.85	-0.48
J111	0.13	-3.68	1.15
J112	0.53	-1.63	-0.08
J113	0.93	0.42	-1.51
J114	0.98	0.67	-2.17
J115	0.54	-1.58	-0.10
J116	0.99	0.72	-2.58
J117	0.80	-0.24	-0.86
J118	0.70	-0.75	-0.54
J119	0.90	0.26	-1.31
K120	0.95	0.49	-1.64
K121	0.85	-0.01	-1.04
K122	0.57	-1.43	-0.18
K123	0.61	-1.20	-0.29
K124	0.76	-0.47	-0.71
K125	0.82	-0.17	-0.92
K126	0.76	-0.44	-0.72
K127	0.98	0.64	-2.05
K128	0.68	-0.85	-0.48
K129	0.88	0.11	-1.15
L130	0.97	0.62	-1.96
L131	0.75	-0.49	-0.69
L132	0.99	0.72	-2.58
L133	0.32	-2.69	0.47
L134	0.68	-0.85	-0.48
L135	0.96	0.54	-1.75
L136	0.28	-2.87	0.57
L137	0.97	0.59	-1.88
L138	0.42	-2.16	0.19
L139	0.29	-2.84	0.55
L140	0.92	0.36	-1.44
L141	0.76	-0.44	-0.72

No. of persons = 200

C.6 Cloze-Type Test (Low Scorers, Malaysia): Facility Values, Item Z-Scores & Item Z-Scale Values

ITEM NAME	FACILITY VALUE	ITEM Z-SCORE	ITEM Z-SCALE VALUE
A 1	0.85	2.05	-1.06
A 2	0.80	1.86	-0.86
A 3	0.73	1.57	-0.61
A 4	0.04	-1.08	1.70
A 5	0.08	-0.95	1.41
A 6	0.56	0.93	-0.16
A 7	0.37	0.18	0.33
A 8	0.75	1.65	-0.67
A 9	0.78	1.78	-0.79
A 10	0.82	1.92	-0.92
A 11	0.73	1.59	-0.63
B 12	0.55	0.87	-0.13
B 13	0.75	1.65	-0.67
B 14	0.32	-0.02	0.47
B 15	0.30	-0.10	0.52
B 16	0.17	-0.60	0.95
B 17	0.24	-0.31	0.69
B 18	0.73	1.59	-0.63
B 19	0.58	0.99	-0.20
B 20	0.11	-0.81	1.20
B 21	0.59	1.03	-0.23
B 22	0.66	1.32	-0.43
B 23	0.61	1.10	-0.28
B 24	0.49	0.64	0.03
C 25	0.88	2.17	-1.20
C 26	0.56	0.91	-0.15
C 27	0.07	-0.97	1.44
C 28	0.02	-1.16	1.96
C 29	0.23	-0.37	0.74
C 30	0.56	0.93	-0.16
C 31	0.80	1.84	-0.84
C 32	0.67	1.34	-0.44
C 33	0.83	1.98	-0.97
C 34	0.70	1.45	-0.52
C 35	0.33	0.02	0.44
C 36	0.40	0.29	0.25
C 37	0.27	-0.19	0.60
D 38	0.63	1.20	-0.35
D 39	0.41	0.35	0.21
D 40	0.63	1.20	-0.35
D 41	0.71	1.51	-0.57
D 42	0.31	-0.06	0.50
D 43	0.43	0.41	0.18
D 44	0.62	1.14	-0.31
D 45	0.51	0.74	-0.04
D 46	0.16	-0.62	0.97
D 47	0.65	1.28	-0.40
D 48	0.47	0.56	0.08
D 49	0.29	-0.13	0.55
D 50	0.07	-0.99	1.48

ITEM NAME	FACILITY VALUE	ITEM Z-SCORE	ITEM Z-SCALE VALUE
D 51	0.57	0.95	-0.18
D 52	0.65	1.28	-0.40
E 53	0.31	-0.06	0.50
E 54	0.17	-0.60	0.95
E 55	0.54	0.83	-0.10
E 56	0.49	0.64	0.03
E 57	0.57	0.95	-0.18
E 58	0.52	0.78	-0.06
E 59	0.04	-1.10	1.75
E 60	0.61	1.12	-0.29
E 61	0.22	-0.39	0.76
E 62	0.07	-0.99	1.48
E 63	0.25	-0.27	0.66
E 64	0.04	-1.08	1.70
F 65	0.57	0.95	-0.18
F 66	0.11	-0.83	1.23
F 67	0.52	0.76	-0.05
F 68	0.19	-0.50	0.86
F 69	0.12	-0.79	1.18
F 70	0.36	0.16	0.35
F 71	0.51	0.72	-0.03
F 72	0.41	0.35	0.21
F 73	0.70	1.47	-0.54
F 74	0.21	-0.44	0.81
F 75	0.19	-0.52	0.88
F 76	0.08	-0.95	1.41
G 77	0.10	-0.85	1.25
G 78	0.03	-1.12	1.81
G 79	0.07	-0.97	1.44
G 80	0.49	0.66	0.01
G 81	0.64	1.22	-0.36
G 82	0.06	-1.01	1.51
G 83	0.50	0.68	0.00
G 84	0.64	1.24	-0.37
G 85	0.56	0.93	-0.16
G 86	0.00	-1.26	4.00
G 87	0.24	-0.33	0.71
H 88	0.52	0.78	-0.06
H 89	0.03	-1.14	1.88
H 90	0.35	0.10	0.39
H 91	0.03	-1.12	1.81
H 92	0.03	-1.12	1.81
H 93	0.50	0.70	-0.01
H 94	0.13	-0.75	1.13
H 95	0.03	-1.14	1.88
H 96	0.21	-0.42	0.79
H 97	0.22	-0.39	0.76
I 98	0.01	-1.22	2.33
I 99	0.38	0.19	0.32
I100	0.01	-1.20	2.17

ITEM NAME	FACILITY VALUE	ITEM Z-SCORE	ITEM Z-SCALE VALUE
I101	0.04	-1.10	1.75
I102	0.31	-0.06	0.50
I103	0.18	-0.56	0.92
I104	0.01	-1.20	2.17
I105	0.05	-1.04	1.60
I106	0.11	-0.83	1.23
I107	0.31	-0.04	0.48
I108	0.29	-0.13	0.55
J109	0.00	-1.26	4.00
J110	0.02	-1.18	2.05
J111	0.02	-1.18	2.05
J112	0.05	-1.04	1.60
J113	0.41	0.35	0.21
J114	0.59	1.03	-0.23
J115	0.05	-1.04	1.60
J116	0.11	-0.83	1.23
J117	0.22	-0.41	0.77
J118	0.03	-1.12	1.81
J119	0.20	-0.46	0.82
K120	0.41	0.35	0.21
K121	0.18	-0.54	0.90
K122	0.00	-1.26	4.00
K123	0.13	-0.77	1.15
K124	0.00	-1.26	4.00
K125	0.17	-0.60	0.95
K126	0.05	-1.04	1.60
K127	0.24	-0.31	0.69
K128	0.01	-1.20	2.17
K129	0.13	-0.75	1.13
L130	0.43	0.43	0.16
L131	0.00	-1.26	4.00
L132	0.18	-0.54	0.90
L133	0.01	-1.24	2.58
L134	0.00	-1.26	4.00
L135	0.16	-0.62	0.97
L136	0.03	-1.14	1.88
L137	0.09	-0.91	1.34
L138	0.01	-1.24	2.58
L139	0.00	-1.26	4.00
L140	0.40	0.29	0.25
L141	0.25	-0.29	0.67

No. of persons = 200

APPENDIX D
TRADITIONAL STATISTICS FOR CLOZE-TYPE TEST (TANZANIAN DATA)

D.1 Cloze-Type Test (Tanzanian Group): Raw Score Distribution & Frequency Counts, K-R20 & SEM

RAW SCORE	FREQ.	CUM. FREQ.	NO. OF PERSONS 0 10 20
0	0	0	|
1	0	0	|
2	0	0	|
3	1	1	|*
4	5	6	|*****
5	0	6	|
6	1	7	|*
7	1	8	|*
8	4	12	|****
9	0	12	|
10	1	13	|*
11	5	18	|*****
12	3	21	|***
13	0	21	|
14	3	24	|***
15	3	27	|***
16	1	28	|*
17	2	30	|**
18	3	33	|***
19	1	34	|*
20	2	36	|**
21	1	37	|*
22	4	41	|****
23	4	45	|****
24	3	48	|***
25	2	50	|**
26	4	54	|****
27	2	56	|**
28	0	56	|
29	1	57	|*
30	0	57	|
31	4	61	|****
32	0	61	|
33	3	64	|***
34	3	67	|***
35	3	70	|***
36	1	71	|*
37	4	75	|****
38	2	77	|**
39	1	78	|*
40	2	80	|**
41	1	81	|*
42	1	82	|*
43	3	85	|***
44	1	86	|*
45	4	90	|****
46	3	93	|***
47	2	95	|**
48	0	95	|
49	5	100	|*****
50	1	101	|*

RAW SCORE	FREQ.	CUM. FREQ.	NO. OF PERSONS 0 10 20
51	2	103	**
52	1	104	*
53	3	107	***
54	3	110	***
55	0	110	
56	3	113	***
57	4	117	****
58	3	120	***
59	1	121	*
60	1	122	*
61	2	124	**
62	3	127	***
63	2	129	**
64	3	132	***
65	2	134	**
66	1	135	*
67	2	137	**
68	2	139	**
69	5	144	*****
70	0	144	
71	4	148	****
72	0	148	
73	1	149	*
74	5	154	*****
75	3	157	***
76	2	159	**
77	4	163	****
78	4	167	****
79	0	167	
80	4	171	****
81	2	173	**
82	3	176	***
83	1	177	*
84	3	180	***
85	2	182	**
86	2	184	**
87	2	186	**
88	2	188	**
89	1	189	*
90	1	190	*
91	4	194	****
92	1	195	*
93	1	196	*
94	3	199	***
95	0	199	
96	3	202	***
97	4	206	****
98	2	208	**
99	3	211	***
100	3	214	***

RAW SCORE	FREQ.	CUM. FREQ.	NO. OF PERSONS 0 10 20 | | |
101	1	215	|*
102	3	218	|***
103	2	220	|**
104	1	221	|*
105	2	223	|**
106	2	225	|**
107	3	228	|***
108	2	230	|**
109	3	233	|***
110	3	236	|***
111	1	237	|*
112	2	239	|**
113	0	239	|
114	0	239	|
115	0	239	|
116	1	240	|*
117	0	240	|
118	1	241	|*
119	1	242	|*
120	0	242	|
121	0	242	|
122	0	242	|
123	0	242	|
124	0	242	|
125	0	242	|
126	0	242	|
127	0	242	|
128	0	242	|
129	1	243	|*
130	0	243	|
131	0	243	|
132	0	243	|
133	0	243	|
134	0	243	|
135	0	243	|
136	0	243	|
137	0	243	|
138	0	243	|
139	0	243	|
140	0	243	|
141	0	243	|

Mean raw score = 59.30
Standard deviation = 32.16
Raw score range for group: 3 to 129

K-R20 = 0.98
SEM = 4.43

D.2 Cloze-Type Test (Tanzanian Group): Traditional Item Statistics

ITEM LABEL	FACILITY VALUE	DISCRIM INDEX	UNBIASED PT.BISERIAL
A 1	0.91	0.27	0.36
A 2	0.77	0.73	0.66
A 3	0.72	0.71	0.59
A 4	0.03	0.08	0.20
A 5	0.31	0.45	0.41
A 6	0.57	0.61	0.52
A 7	0.25	0.27	0.26
A 8	0.72	0.32	0.24
A 9	0.81	0.38	0.40
A 10	0.85	0.45	0.49
A 11	0.86	0.39	0.47
B 12	0.63	0.86	0.69
B 13	0.76	-0.02	-0.01
B 14	0.55	0.82	0.62
B 15	0.49	0.95	0.76
B 16	0.26	0.53	0.45
B 17	0.37	0.64	0.53
B 18	0.78	0.47	0.47
B 19	0.67	0.61	0.50
B 20	0.11	0.26	0.34
B 21	0.58	0.86	0.69
B 22	0.73	0.42	0.39
B 23	0.68	0.85	0.70
B 24	0.60	0.61	0.40
C 25	0.88	0.20	0.25
C 26	0.62	0.88	0.71
C 27	0.43	0.77	0.65
C 28	0.25	0.50	0.44
C 29	0.53	0.89	0.71
C 30	0.69	0.83	0.68
C 31	0.81	0.36	0.37
C 32	0.56	0.68	0.51
C 33	0.66	0.82	0.62
C 34	0.51	0.79	0.58
C 35	0.44	0.64	0.52
C 36	0.67	0.71	0.63
C 37	0.53	0.89	0.70
D 38	0.62	0.74	0.62
D 39	0.57	0.62	0.43
D 40	0.68	0.82	0.68
D 41	0.67	0.50	0.38
D 42	0.30	0.68	0.59
D 43	0.71	0.50	0.46
D 44	0.47	0.71	0.57
D 45	0.56	0.76	0.61
D 46	0.21	0.45	0.44
D 47	0.59	0.74	0.57
D 48	0.49	0.27	0.20
D 49	0.56	0.74	0.55
D 50	0.24	0.55	0.48

ITEM LABEL	FACILITY VALUE	DISCRIM INDEX	UNBIASED PT.BISERIAL
D 51	0.60	0.65	0.53
D 52	0.61	0.39	0.28
E 53	0.35	0.76	0.62
E 54	0.21	0.55	0.50
E 55	0.63	0.86	0.72
E 56	0.72	0.71	0.60
E 57	0.53	0.91	0.72
E 58	0.55	0.79	0.62
E 59	0.78	0.36	0.37
E 60	0.65	0.58	0.50
E 61	0.30	0.62	0.55
E 62	0.17	0.55	0.53
E 63	0.53	0.82	0.66
E 64	0.11	0.35	0.43
F 65	0.74	0.47	0.44
F 66	0.46	0.98	0.76
F 67	0.51	0.95	0.73
F 68	0.48	0.67	0.56
F 69	0.40	0.80	0.64
F 70	0.58	0.70	0.57
F 71	0.44	0.77	0.57
F 72	0.58	0.85	0.67
F 73	0.72	0.71	0.60
F 74	0.28	0.44	0.33
F 75	0.35	0.85	0.69
F 76	0.41	0.73	0.60
G 77	0.19	0.32	0.35
G 78	0.19	0.55	0.54
G 79	0.19	0.48	0.46
G 80	0.63	0.83	0 69
G 81	0.81	0.55	0.57
G 82	0.19	0.55	0.53
G 83	0.73	0.73	0.62
G 84	0.58	0.56	0.47
G 85	0.67	0.76	0.65
G 86	0.31	0.73	0.61
G 87	0.36	0.92	0.74
H 88	0.70	0.73	0.62
H 89	0.21	0.52	0.48
H 90	0.53	0.89	0.71
H 91	0.15	0.35	0.37
H 92	0.03	0.06	0.14
H 93	0.65	0.85	0.70
H 94	0.15	0.24	0.27
H 95	0.07	0.20	0.30
H 96	0.27	0.70	0.60
H 97	0.18	0.45	0.48
I 98	0.00	0.00	0.04
I 99	0.47	0.77	0.58

ITEM LABEL	FACILITY VALUE	DISCRIM INDEX	UNBIASED PT.BISERIAL
I100	0.16	0.50	0.52
I101	0.09	0.18	0.26
I102	0.45	0.86	0.69
I103	0.51	0.85	0.66
I104	0.28	0.65	0.58
I105	0.12	0.33	0.42
I106	0.16	0.47	0.51
I107	0.37	0.71	0.56
I108	0.39	0.76	0.61
J109	0.00	0.02	0.10
J110	0.14	0.41	0.46
J111	0.01	0.02	0.06
J112	0.14	0.36	0.40
J113	0.51	0.83	0.61
J114	0.68	0.88	0.70
J115	0.20	0.38	0.38
J116	0.40	0.89	0.73
J117	0.24	0.38	0.38
J118	0.17	0.41	0.41
J119	0.36	0.53	0.43
K120	0.37	0.59	0.45
K121	0.28	0.61	0.52
K122	0.04	0.08	0.14
K123	0.21	0.41	0.38
K124	0.15	0.42	0.47
K125	0.21	0.42	0.43
K126	0.28	0.79	0.67
K127	0.33	0.68	0.56
K128	0.14	0.39	0.43
K129	0.24	0.53	0.49
L130	0.03	0.05	0.10
L131	0.18	0.52	0.53
L132	0.48	0.94	0.73
L133	0.11	0.27	0.37
L134	0.13	0.39	0.46
L135	0.33	0.77	0.63
L136	0.09	0.18	0.23
L137	0.26	0.71	0.63
L138	0.05	0.17	0.30
L139	0.06	0.21	0.36
L140	0.46	0.92	0.72
L141	0.28	0.48	0.43

No. of persons = 243

D.3 Grouped Item Statistics (Tanzanian Data)

Table 1 Cloze-Type Test (Tanzanian Data): Items Grouped by Facility Value

Facility Value Interval	No.of Items	Item Names
0.90-1.00	1	A1
0.80-0.89	6	A9 A10 A11 C25 C31 G81
0.70-0.79	13	A2 A3 A8 B13 B18 B22 D43 E56 E59 F65 F73 G83 H88
0.60-0.69	19	B12 B19 B23 B24 C26 C30 C33 C36 D38 D40 D41 D51 D52 E55 E60 G80 D52 E55 E60 G80
0.50-0.59	21	A6 B14 B21 C29 C32 C34 C37 D39 D45 D47 D49 E57 E58 E63 F67 F70 F72 G84 H90 I103 J113
0.40-0.49	15	B15 C27 C35 D44 D48 F66 F68 F69 F71 F76 I99 I102 J116 L132 L140
0.30-0.39	14	A5 B17 D42 E53 E61 F75 G86 G87 I107 I108 J119 K120 K127 L135
0.20-0.29	19	A7 B16 C28 D46 D50 E54 F74 H89 H96 I104 J115 J117 K121 K123 K125 K126 K129 L137 L141
0.10-0.19	21	B20 E62 E64 G77 G78 G79 G82 H91 H94 H97 I100 I105 I106 J110 J112 J118 K124 K128 L131 L133 L134
0.00-0.09	12	A4 H92 H95 I98 I101 J109 J111 K122 L130 L136 L138 L139

Facility value range = 0.00 (items I98,J109) to 0.91 (item A1)
Mean = 0.42
SD = 0.24

Table 2 Cloze-Type Test (Tanzanian Data): Items Grouped by Discrimination Index

Discrim. Index Interval	No.of Items	Item Names
0.90-1.00	7	B15 E57 F66 F67 G87 L132 L140
0.80-0.89	23	B12 B14 B21 B23 C26 C29 C30 C33 C37 D40 E55 E63 F69 F72 F75 G80 H90 H93 I102 I103 J113 J114 J116
0.70-0.79	28	A2 A3 C27 C34 C36 D38 D44 D45 D47 D49 E53 E56 E58 F70 F71 F73 F76 G83 G85 G86 H88 H96 I99 I107 I108 K126 L135 L137
0.60-0.69	14	A6 B17 B19 B24 C32 C35 D39 D42 D51 E61 F68 I104 K121 K127
0.50-0.59	18	B16 C28 D41 D43 D50 E54 E60 E62 G78 G81 G82 G84 H89 I100 J119 K120 K129 L131
0.40-0.49	16	A5 A10 B18 B22 D46 F65 F74 G79 H97 I106 J110 J118 K123 K124 K125 L141
0.30-0.39	15	A8 A9 A11 C31 D52 E59 E64 G77 H91 I105 J112 J115 J117 K128 L134
0.20-0.29	9	A1 A7 B20 C25 D48 H94 H95 L133 L139
0.10-0.19	3	I101 L136 L138
0.00-0.09	7	A4 H92 I98 J109 J111 K122 L130
-0.09- -0.01	1	B13

Discrimination index range =-0.02 (Item B13) to 0.98 (Item F66)
Mean = 0.57
SD = 0.24

Table 3 Cloze-Type Test (Tanzanian Data): Items Grouped by Unbiased Point Biserial

Point Biserial Interval	No.of Items	Item Names
0.90-1.00	0	
0.80-0.89	0	
0.70-0.79	16	B15 B23 C26 C29 C37 E55 E57 F66 F67 G87 H90 H93 J114 J116 L132 L140
0.60-0.69	33	A2 B12 B14 B21 C27 C30 C33 C36 D38 D40 D45 E53 E56 E58 E63 F69 F72 F73 F75 F76 G80 G83 G85 G86 H88 H96 I102 I103 I108 J113 K126 L135 L137
0.50-0.59	30	A3 A6 B17 B19 C32 C34 C35 D42 D44 D47 D49 D51 E54 E60 E61 E62 F68 F70 F71 G78 G81 G82 I99 I100 I104 I106 I107 K121 K127 L131
0.40-0.49	30	A5 A9 A10 A11 B16 B18 B24 C28 D39 D43 D46 D50 E64 F65 G79 G84 H89 H97 I105 J110 J112 J118 J119 K120 K124 K125 K128 K129 L134 L141
0.30-0.39	16	A1 B20 B22 C31 D41 E59 F74 G77 H91 H95 J115 J117 K123 L133 L138 L139
0.20-0.29	9	L136 A4 A7 A8 C25 D48 D52 H94 I101
0.10-0.19	4	H92 J109 K122 L130
0.00-0.09	2	I98 J111
-0.09- -0.01	1	B13

Point biserial range =-0.01 (Item B13) to 0.76 (Items B15,F66)
Mean = 0.5
SD = 0.17

D.4 Cloze-Type Test (Tanzanian Group): Item Z-Scores & Item Z-Scale Values

ITEM NAME	ITEM Z-SCORE	ITEM Z-SCALE VALUE
A 1	2.04	-1.31
A 2	1.47	-0.74
A 3	1.24	-0.57
A 4	-1.65	1.90
A 5	-0.45	0.49
A 6	0.64	-0.18
A 7	-0.73	0.68
A 8	1.28	-0.59
A 9	1.66	-0.90
A 10	1.80	-1.03
A 11	1.86	-1.10
B 12	0.88	-0.33
B 13	1.41	-0.70
B 14	0.55	-0.13
B 15	0.29	0.30
B 16	-0.70	0.66
B 17	-0.20	0.32
B 18	1.52	-0.78
B 19	1.05	-0.44
B 20	-1.32	1.24
B 21	0.65	-0.19
B 22	1.31	-0.62
B 23	1.10	-0.48
B 24	0.74	-0.24
C 25	1.94	-1.18
C 26	0.84	-0.31
C 27	0.05	0.17
C 28	-0.71	0.67
C 29	0.48	-0.09
C 30	1.12	-0.49
C 31	1.64	-0.88
C 32	0.60	-0.16
C 33	1.02	-0.42
C 34	0.39	-0.03
C 35	0.06	0.16
C 36	1.07	-0.45
C 37	0.45	-0.07
D 38	0.84	-0.31
D 39	0.64	-0.18
D 40	1.09	-0.46
D 41	1.04	-0.43
D 42	-0.52	0.54
D 43	1.22	-0.56
D 44	0.20	0.08
D 45	0.60	-0.16
D 46	-0.87	0.79
D 47	0.70	-0.22
D 48	0.28	0.03
D 49	0.57	-0.14
D 50	-0.76	0.71

ITEM NAME	ITEM Z-SCORE	ITEM Z-SCALE VALUE
D 51	0.74	-0.24
D 52	0.79	-0.28
E 53	-0.28	0.37
E 54	-0.87	0.79
E 55	0.88	-0.33
E 56	1.24	-0.57
E 57	0.46	-0.08
E 58	0.53	-0.12
E 59	1.50	-0.77
E 60	0.96	-0.39
E 61	-0.52	0.54
E 62	-1.06	0.96
E 63	0.46	-0.08
E 64	-1.30	1.22
F 65	1.36	-0.66
F 66	0.15	0.11
F 67	0.36	-0.01
F 68	0.25	0.05
F 69	-0.07	0.24
F 70	0.65	-0.19
F 71	0.10	0.14
F 72	0.67	-0.20
F 73	1.28	-0.60
F 74	-0.57	0.57
F 75	-0.30	0.39
F 76	-0.06	0.23
G 77	-0.96	0.87
G 78	-0.97	0.88
G 79	-0.96	0.87
G 80	0.90	-0.34
G 81	1.62	-0.87
G 82	-0.99	0.90
G 83	1.31	-0.62
G 84	0.67	-0.20
G 85	1.05	-0.44
G 86	-0.47	0.50
G 87	-0.25	0.35
H 88	1.17	-0.52
H 89	-0.87	0.79
H 90	0.46	-0.08
H 91	-1.13	1.03
H 92	-1.65	1.90
H 93	0.98	-0.40
H 94	-1.13	1.03
H 95	-1.47	1.48
H 96	-0.62	0.61
H 97	-1.02	0.93
I 98	-1.75	2.65
I 99	0.19	0.09
I100	-1.11	1.01

ITEM NAME	ITEM Z-SCORE	ITEM Z-SCALE VALUE
I101	-1.41	1.37
I102	0.14	0.12
I103	0.36	-0.01
I104	-0.57	0.57
I105	-1.27	1.18
I106	-1.10	0.99
I107	-0.23	0.34
I108	-0.12	0.28
J109	-1.75	2.65
J110	-1.16	1.06
J111	-1.72	2.26
J112	-1.18	1.08
J113	0.39	-0.03
J114	1.09	-0.46
J115	-0.94	0.85
J116	-0.11	0.27
J117	-0.76	0.71
J118	-1.04	0.94
J119	-0.26	0.36
K120	-0.23	0.34
K121	-0.59	0.58
K122	-1.61	0.33
K123	-0.90	0.82
K124	-1.15	1.05
K125	-0.90	0.82
K126	-0.59	0.58
K127	-0.40	0.45
K128	-1.20	1.10
K129	-0.75	0.70
L130	-1.63	1.84
L131	-1.01	0.91
L132	0.25	0.05
L133	-1.32	1.24
L134	-1.23	1.14
L135	-0.37	0.43
L136	-1.41	1.37
L137	-0.68	0.65
L138	-1.54	1.62
L139	-1.51	1.54
L140	0.15	0.11
L141	-0.59	0.58

No. of persons = 243

APPENDIX E
RASCH STATISTICS FOR CLOZE-TYPE TEST (MALAYSIAN DATA)

E.1 Cloze-Type Test (Malaysian Group): Raw Scores, Rasch Ability Estimates and Standard Errors

TABLE 1
(All measurable persons included)

RAW SCORE	FREQ. COUNT	ABILITY ESTIM.	STANDARD ERROR
140	0	6.49	1.03
139	0	5.75	0.74
138	0	5.29	0.62
137	0	4.96	0.55
136	1	4.69	0.50
135	0	4.46	0.46
134	2	4.26	0.43
133	5	4.08	0.41
132	1	3.93	0.39
131	6	3.78	0.38
130	6	3.64	0.36
129	3	3.52	0.35
128	7	3.40	0.34
127	9	3.29	0.33
126	8	3.18	0.32
125	11	3.08	0.31
124	7	2.99	0.31
123	15	2.89	0.30
122	8	2.81	0.30
121	8	2.72	0.29
120	11	2.64	0.29
119	6	2.56	0.28
118	16	2.48	0.28
117	8	2.41	0.27
116	6	2.33	0.27
115	10	2.26	0.27
114	7	2.19	0.26
113	4	2.12	0.26
112	7	2.06	0.26
111	11	1.99	0.25
110	8	1.93	0.25
109	5	1.86	0.25
108	7	1.80	0.25
107	4	1.74	0.25
106	12	1.68	0.24
105	9	1.62	0.24
104	14	1.57	0.24
103	8	1.51	0.24
102	7	1.45	0.24
101	3	1.40	0.23
100	5	1.34	0.23
99	9	1.29	0.23
98	8	1.24	0.23
97	8	1.18	0.23
96	4	1.13	0.23
95	3	1.08	0.23
94	4	1.03	0.23
93	7	0.98	0.22
92	7	0.93	0.22
91	5	0.88	0.22

TABLE 2
(6 misfitting persons excluded)

RAW SCORE	FREQ. COUNT	ABILITY ESTIM.	STANDARD ERROR
140	0	6.50	1.03
139	0	5.76	0.74
138	0	5.30	0.62
137	0	4.97	0.55
136	1	4.70	0.50
135	0	4.47	0.46
134	2	4.27	0.43
133	5	4.10	0.41
132	1	3.94	0.39
131	6	3.79	0.38
130	6	3.66	0.36
129	3	3.53	0.35
128	7	3.41	0.34
127	9	3.30	0.33
126	8	3.19	0.32
125	11	3.09	0.31
124	7	3.00	0.31
123	15	2.91	0.30
122	8	2.82	0.30
121	8	2.73	0.29
120	11	2.65	0.29
119	6	2.57	0.28
118	16	2.49	0.28
117	8	2.42	0.27
116	6	2.34	0.27
115	10	2.27	0.27
114	7	2.20	0.26
113	4	2.13	0.26
112	7	2.07	0.26
111	11	2.00	0.26
110	8	1.94	0.25
109	5	1.87	0.25
108	7	1.81	0.25
107	4	1.75	0.25
106	12	1.69	0.24
105	9	1.63	0.24
104	14	1.58	0.24
103	8	1.52	0.24
102	6	1.46	0.24
101	3	1.41	0.24
100	5	1.35	0.23
99	9	1.30	0.23
98	8	1.25	0.23
97	7	1.19	0.23
96	4	1.14	0.23
95	3	1.09	0.23
94	4	1.04	0.23
93	7	0.99	0.23
92	7	0.94	0.22
91	5	0.89	0.22

TABLE 1
(All measurable persons included)

RAW SCORE	FREQ. COUNT	ABILITY ESTIM.	STANDARD ERROR
90	2	0.83	0.22
89	2	0.78	0.22
88	3	0.73	0.22
87	5	0.69	0.22
86	6	0.64	0.22
85	9	0.59	0.22
84	9	0.54	0.22
83	7	0.50	0.22
82	8	0.45	0.22
81	4	0.40	0.22
80	6	0.36	0.22
79	5	0.31	0.21
78	11	0.27	0.21
77	3	0.22	0.21
76	5	0.18	0.21
75	5	0.13	0.21
74	6	0.09	0.21
73	2	0.04	0.21
72	5	-0.00	0.21
71	7	-0.05	0.21
70	4	-0.09	0.21
69	4	-0.14	0.21
68	6	-0.18	0.21
67	6	-0.23	0.21
66	8	-0.27	0.21
65	4	-0.32	0.21
64	1	-0.36	0.21
63	6	-0.40	0.21
62	2	-0.45	0.21
61	7	-0.49	0.21
60	3	-0.54	0.21
59	2	-0.58	0.21
58	4	-0.63	0.21
57	1	-0.67	0.21
56	1	-0.72	0.21
55	6	-0.76	0.21
54	3	-0.81	0.21
53	2	-0.85	0.21
52	4	-0.90	0.21
51	1	-0.94	0.22
50	4	-0.99	0.22
49	3	-1.04	0.22
48	4	-1.08	0.22
47	3	-1.13	0.22
46	5	-1.18	0.22
45	2	-1.23	0.22
44	1	-1.27	0.22
43	2	-1.32	0.22
42	1	-1.37	0.22
41	3	-1.42	0.22

TABLE 2
(6 misfitting persons excluded)

RAW SCORE	FREQ. COUNT	ABILITY ESTIM.	STANDARD ERROR
90	2	0.84	0.22
89	2	0.79	0.22
88	3	0.74	0.22
87	5	0.69	0.22
86	6	0.64	0.22
85	9	0.60	0.22
84	9	0.55	0.22
83	6	0.50	0.22
82	8	0.46	0.22
81	4	0.41	0.22
80	6	0.36	0.22
79	5	0.32	0.22
78	11	0.27	0.21
77	3	0.23	0.21
76	5	0.18	0.21
75	5	0.13	0.21
74	6	0.09	0.21
73	2	0.04	0.21
72	5	-0.00	0.21
71	7	-0.05	0.21
70	4	-0.09	0.21
69	4	-0.13	0.21
68	6	-0.18	0.21
67	6	-0.22	0.21
66	8	-0.27	0.21
65	4	-0.31	0.21
64	1	-0.36	0.21
63	6	-0.40	0.21
62	2	-0.45	0.21
61	7	-0.49	0.21
60	3	-0.54	0.21
59	1	-0.58	0.21
58	4	-0.63	0.21
57	1	-0.67	0.21
56	1	-0.72	0.21
55	6	-0.76	0.21
54	3	-0.81	0.21
53	2	-0.85	0.21
52	4	-0.90	0.22
51	1	-0.95	0.22
50	4	-0.99	0.22
49	3	-1.04	0.22
48	4	-1.09	0.22
47	3	-1.13	0.22
46	5	-1.18	0.22
45	2	-1.23	0.22
44	1	-1.28	0.22
43	2	-1.33	0.22
42	1	-1.37	0.22
41	3	-1.42	0.22

TABLE 1
(All measurable persons included)

RAW SCORE	FREQ. COUNT	ABILITY ESTIM.	STANDARD ERROR
40	2	-1.47	0.22
39	6	-1.52	0.23
38	2	-1.57	0.23
37	1	-1.62	0.23
36	2	-1.67	0.23
35	3	-1.73	0.23
34	3	-1.78	0.23
33	1	-1.84	0.24
32	4	-1.89	0.24
31	2	-1.95	0.24
30	1	-2.00	0.24
29	2	-2.06	0.24
28	2	-2.12	0.25
27	1	-2.18	0.25
26	4	-2.25	0.25
25	1	-2.31	0.25
24	6	-2.38	0.26
23	3	-2.44	0.26
22	2	-2.51	0.27
21	2	-2.58	0.27
20	0	-2.66	0.27
19	2	-2.73	0.28
18	1	-2.81	0.28
17	4	-2.89	0.29
16	1	-2.98	0.30
15	2	-3.07	0.30
14	2	-3.16	0.31
13	3	-3.26	0.32
12	1	-3.37	0.33
11	2	-3.48	0.34
10	1	-3.60	0.36
9	0	-3.73	0.37
8	2	-3.88	0.39
7	1	-4.04	0.41
6	1	-4.22	0.44
5	1	-4.43	0.48
4	0	-4.68	0.53
3	0	-4.99	0.60
2	0	-5.42	0.73
1	1	-6.14	1.01

No. of persons = 608
Mean ability = 0.82
SD ability = 1.78
Group ability range: -6.14 to 4.69

Person separability index = 0.98
No. of person strata = 9.46

TABLE 2
(6 misfitting persons excluded)

RAW SCORE	FREQ. COUNT	ABILITY ESTIM.	STANDARD ERROR
40	2	-1.47	0.23
39	6	-1.53	0.23
38	2	-1.58	0.23
37	1	-1.63	0.23
36	2	-1.68	0.23
35	3	-1.73	0.23
34	3	-1.79	0.23
33	1	-1.84	0.24
32	4	-1.90	0.24
31	2	-1.96	0.24
30	1	-2.01	0.24
29	2	-2.07	0.24
28	1	-2.13	0.25
27	1	-2.19	0.25
26	4	-2.26	0.25
25	1	-2.32	0.26
24	5	-2.39	0.26
23	3	-2.45	0.26
22	2	-2.52	0.27
21	2	-2.60	0.27
20	0	-2.67	0.28
19	2	-2.75	0.28
18	1	-2.83	0.29
17	4	-2.91	0.29
16	1	-2.99	0.30
15	2	-3.08	0.30
14	2	-3.18	0.31
13	3	-3.28	0.32
12	1	-3.39	0.33
11	2	-3.50	0.34
10	1	-3.62	0.36
9	0	-3.75	0.37
8	2	-3.90	0.39
7	1	-4.06	0.41
6	1	-4.24	0.44
5	1	-4.45	0.48
4	0	-4.70	0.53
3	0	-5.02	0.60
2	0	-5.45	0.73
1	1	-6.17	1.01

No. of persons = 602
Mean ability = 0.83
SD ability = 1.79
Group ability range: -6.17 to 4.70

Person separability index = 0.98
No. of person strata = 9.48

E.2 Cloze-Type Test (Malaysian Group): Item Difficulty Estimates & Standard Errors

SET 1 (All measurable persons included)			SET 2 (6 misfitting persons excluded)		
ITEM NAME	ITEM DIFFICULTY	STANDARD ERROR		ITEM DIFFICULTY	STANDARD ERROR
A 1	-3.33	0.20		-3.41	0.21
A 2	-3.27	0.20		-3.33	0.20
A 3	-2.53	0.16		-2.59	0.17
A 4	1.78	0.10		1.78	0.10
A 5	1.18	0.10		1.20	0.10
A 6	-1.01	0.12		-1.03	0.12
A 7	0.52	0.10		0.54	0.10
A 8	-2.59	0.17		-2.62	0.17
A 9	-3.01	0.18		-3.10	0.19
A 10	-3.35	0.20		-3.37	0.20
A 11	-2.59	0.17		-2.59	0.17
B 12	-1.75	0.14		-1.78	0.14
B 13	-1.01	0.12		-1.03	0.12
B 14	-0.14	0.11		-0.11	0.11
B 15	-0.88	0.12		-0.89	0.12
B 16	0.85	0.10		0.85	0.10
B 17	0.88	0.10		0.88	0.10
B 18	-2.33	0.16		-2.38	0.16
B 19	-1.70	0.14		-1.72	0.14
B 20	0.78	0.10		0.80	0.10
B 21	-1.99	0.14		-2.03	0.15
B 22	-2.14	0.15		-2.18	0.15
B 23	-2.08	0.15		-2.11	0.15
B 24	-1.04	0.12		-1.03	0.12
C 25	-3.77	0.23		-3.86	0.24
C 26	-1.75	0.14		-1.80	0.14
C 27	0.83	0.10		0.83	0.10
C 28	2.64	0.11		2.64	0.11
C 29	-0.43	0.11		-0.43	0.11
C 30	-1.75	0.14		-1.76	0.14
C 31	-2.73	0.17		-2.73	0.17
C 32	-1.79	0.14		-1.84	0.14
C 33	-3.43	0.21		-3.46	0.21
C 34	-1.38	0.13		-1.40	0.13
C 35	-0.11	0.11		-0.09	0.11
C 36	-0.14	0.11		-0.13	0.11
C 37	-0.30	0.11		-0.32	0.11
D 38	-2.10	0.15		-2.14	0.15
D 39	-0.77	0.12		-0.79	0.12
D 40	-1.35	0.13		-1.33	0.13
D 41	-2.43	0.16		-2.45	0.16
D 42	-0.15	0.11		-0.15	0.11
D 43	-0.90	0.12		-0.89	0.12
D 44	-1.11	0.12		-1.12	0.12
D 45	-1.28	0.13		-1.27	0.13
D 46	-0.11	0.11		-0.11	0.11
D 47	-1.87	0.14		-1.88	0.14
D 48	-0.57	0.11		-0.56	0.11
D 49	-0.07	0.11		-0.04	0.11
D 50	0.71	0.10		0.72	0.10

SET 1 (All measurable persons included)			SET 2 (6 misfitting persons excluded)		
ITEM NAME	ITEM DIFFICULTY	STANDARD ERROR		ITEM DIFFICULTY	STANDARD ERROR
D 51	-1.85	0.14		-1.84	0.14
D 52	-2.06	0.15		-2.05	0.15
E 53	-0.42	0.11		-0.43	0.11
E 54	1.58	0.10		1.60	0.10
E 55	-1.13	0.12		-1.12	0.12
E 56	-1.47	0.13		-1.45	0.13
E 57	-1.62	0.13		-1.63	0.14
E 58	-1.52	0.13		-1.56	0.13
E 59	2.18	0.11		2.21	0.11
E 60	-1.75	0.14		-1.76	0.14
E 61	0.30	0.11		0.31	0.11
E 62	1.19	0.10		1.20	0.10
E 63	-0.07	0.11		-0.07	0.11
E 64	1.51	0.10		1.52	0.10
F 65	-0.68	0.12		-0.67	0.12
F 66	0.88	0.10		0.88	0.10
F 67	-0.57	0.11		-0.59	0.11
F 68	0.64	0.10		0.62	0.10
F 69	0.67	0.10		0.67	0.10
F 70	-0.33	0.11		-0.33	0.11
F 71	-1.24	0.13		-1.25	0.13
F 72	-0.76	0.12		-0.79	0.12
F 73	-1.48	0.13		-1.45	0.13
F 74	0.22	0.11		0.23	0.11
F 75	-0.07	0.11		-0.06	0.11
F 76	1.23	0.10		1.24	0.10
G 77	1.08	0.10		1.10	0.10
G 78	1.78	0.10		1.81	0.10
G 79	1.29	0.10		1.31	0.10
G 80	-1.33	0.13		-1.32	0.13
G 81	-1.47	0.13		-1.48	0.13
G 82	2.20	0.11		2.22	0.11
G 83	-1.48	0.13		-1.50	0.13
G 84	-2.06	0.15		-2.07	0.15
G 85	-1.75	0.14		-1.76	0.14
G 86	1.72	0.10		1.72	0.10
G 87	-0.23	0.11		-0.22	0.11
H 88	-1.24	0.13		-1.23	0.13
H 89	3.18	0.12		3.18	0.12
H 90	-0.12	0.11		-0.11	0.11
H 91	1.26	0.10		1.27	0.10
H 92	2.90	0.12		2.93	0.12
H 93	-1.47	0.13		-1.45	0.13
H 94	1.53	0.10		1.53	0.10
H 95	2.79	0.11		2.81	0.11
H 96	-0.19	0.11		-0.19	0.11
H 97	0.25	0.11		0.25	0.11
I 98	4.26	0.16		4.26	0.16
I 99	-0.69	0.12		-0.67	0.12
I100	1.08	0.10		1.09	0.10

SET 1 (All measurable persons included)			SET 2 (6 misfitting persons excluded)	
ITEM NAME	ITEM DIFFICULTY	STANDARD ERROR	ITEM DIFFICULTY	STANDARD ERROR
I101	3.10	0.12	3.11	0.12
I102	-0.62	0.12	-0.61	0.12
I103	0.04	0.11	0.06	0.11
I104	2.09	0.11	2.12	0.11
I105	1.04	0.10	1.07	0.10
I106	0.72	0.10	0.73	0.10
I107	-0.83	0.12	-0.80	0.12
I108	-0.53	0.11	-0.53	0.11
J109	4.16	0.16	4.16	0.16
J110	2.16	0.11	2.17	0.11
J111	4.78	0.19	4.79	0.19
J112	2.36	0.11	2.39	0.11
J113	-0.57	0.11	-0.56	0.11
J114	-1.97	0.14	-1.96	0.14
J115	2.29	0.11	2.33	0.11
J116	0.02	0.11	0.04	0.11
J117	0.55	0.10	0.56	0.10
J118	1.85	0.10	1.87	0.10
J119	0.25	0.11	0.27	0.11
K120	-0.50	0.11	-0.48	0.11
K121	0.83	0.10	0.82	0.10
K122	2.79	0.11	2.79	0.11
K123	1.81	0.10	1.82	0.10
K124	2.10	0.11	2.11	0.11
K125	0.86	0.10	0.85	0.10
K126	1.59	0.10	1.60	0.10
K127	-0.36	0.11	-0.34	0.11
K128	2.16	0.11	2.20	0.11
K129	0.68	0.10	0.68	0.10
L130	-0.97	0.12	-0.96	0.12
L131	2.11	0.11	2.11	0.11
L132	-0.26	0.11	-0.27	0.11
L133	3.77	0.14	3.77	0.14
L134	2.44	0.11	2.46	0.11
L135	0.12	0.11	0.11	0.11
L136	3.08	0.12	3.08	0.12
L137	0.31	0.11	0.32	0.11
L138	3.08	0.12	3.09	0.12
L139	3.75	0.14	3.75	0.14
L140	-0.69	0.12	-0.69	0.12
L141	0.65	0.10	0.65	0.10
Items calibrated on 608 persons Mean item difficulty = 0.00 SD item difficulty = 1.80 Difficulty range: -3.77 to 4.78			Items calibrated on 602 persons Mean item difficulty = 0.00 SD item difficulty = 1.81 Difficulty range: -3.86 to 4.79	

E.3 Cloze-Type Test (Malaysian Group): Observed ICCs & Departures from Expectation

ITEM NAME	ITEM CHARACTERISTIC CURVE SUBGROUP 1	2	3	4	5	6	DEPARTURE FROM EXPECTED ICC SUBGROUP 1	2	3	4	5	6
A 1	0.84	0.92	0.97	0.96	0.99	1.00	0.09	-0.04	-0.01	-0.03	-0.01	0.00
A 2	0.69	0.97	1.00	1.00	1.00	1.00	-0.05	0.02	0.02	0.01	0.00	0.00
A 3	0.58	0.93	0.95	0.99	0.99	1.00	-0.02	0.03	-0.01	0.01	-0.00	0.00
A 4	0.03	0.07	0.13	0.48	0.69	0.82	0.00	-0.04	-0.10	0.05	0.07	0.02
A 5	0.06	0.13	0.25	0.57	0.84	0.92	0.01	-0.05	-0.10	0.01	0.09	0.04
A 6	0.46	0.72	0.73	0.85	0.93	0.98	0.17	0.05	-0.10	-0.07	-0.03	-0.01
A 7	0.39	0.37	0.33	0.46	0.88	0.98	0.30	0.07	-0.18	-0.26	0.03	0.05
A 8	0.66	0.88	0.94	0.99	0.99	0.99	0.05	-0.03	-0.02	0.01	-0.00	-0.01
A 9	0.70	0.93	1.00	0.99	1.00	0.98	0.00	-0.01	0.03	0.00	0.00	-0.02
A 10	0.71	0.96	1.00	1.00	1.00	1.00	-0.04	0.01	0.02	0.01	0.00	0.00
A 11	0.62	0.87	0.95	1.00	1.00	1.00	0.02	-0.03	-0.01	0.02	0.01	0.00
B 12	0.28	0.87	0.97	0.98	1.00	1.00	-0.16	0.06	0.06	0.02	0.02	0.01
B 13	0.78	0.78	0.68	0.75	0.78	0.90	0.49	0.11	-0.15	-0.17	-0.19	-0.09
B 14	0.27	0.38	0.58	0.81	0.93	1.00	0.12	-0.06	-0.08	-0.02	0.01	0.04
B 15	0.05	0.61	0.92	0.98	1.00	1.00	-0.21	-0.03	0.12	0.07	0.04	0.02
B 16	0.10	0.26	0.42	0.76	0.71	0.85	0.03	0.02	-0.01	0.11	-0.10	-0.07
B 17	0.15	0.35	0.40	0.64	0.68	0.85	0.09	0.12	-0.02	0.00	-0.12	-0.06
B 18	0.61	0.90	0.90	0.97	0.98	1.00	0.05	0.02	-0.05	-0.01	-0.01	0.00
B 19	0.45	0.75	0.89	0.98	1.00	1.00	0.02	-0.05	-0.01	0.02	0.02	0.01
B 20	0.08	0.15	0.25	0.73	0.96	0.98	0.01	-0.10	-0.19	0.07	0.15	0.06
B 21	0.39	0.86	0.97	1.00	1.00	1.00	-0.10	0.02	0.04	0.03	0.01	0.01
B 22	0.54	0.84	0.94	0.97	0.99	1.00	0.02	-0.02	0.00	-0.00	0.00	0.00
B 23	0.39	0.88	0.99	1.00	1.00	1.00	-0.12	0.03	0.06	0.03	0.01	0.01
B 24	0.41	0.62	0.77	0.93	0.96	0.98	0.12	-0.05	-0.06	0.01	-0.00	-0.01
C 25	0.83	0.99	0.99	0.98	0.99	0.99	0.01	0.02	0.00	-0.01	-0.01	-0.01
C 26	0.33	0.86	0.92	1.00	1.00	1.00	-0.11	0.05	0.01	0.04	0.02	0.01
C 27	0.00	0.17	0.41	0.69	0.87	0.98	-0.07	-0.07	-0.03	0.04	0.06	0.07
C 28	0.02	0.04	0.11	0.18	0.39	0.72	0.01	-0.01	-0.01	-0.05	-0.02	0.08
C 29	0.06	0.45	0.84	0.89	0.98	0.99	-0.13	-0.07	0.12	0.03	0.04	0.02
C 30	0.37	0.82	0.95	0.98	0.99	0.98	-0.07	0.01	0.04	0.02	0.01	-0.01
C 31	0.75	0.87	0.94	0.96	0.99	0.98	0.12	-0.05	-0.02	-0.02	-0.00	-0.02
C 32	0.54	0.84	0.90	0.89	0.97	0.98	0.09	0.02	-0.01	-0.07	-0.01	-0.01
C 33	0.74	0.97	0.98	1.00	1.00	1.00	-0.02	0.01	-0.00	0.01	0.00	0.00
C 34	0.52	0.93	0.90	0.88	0.79	0.86	0.17	0.19	0.03	-0.06	-0.19	-0.13
C 35	0.17	0.52	0.68	0.81	0.87	0.91	0.02	0.07	0.02	-0.02	-0.04	-0.06
C 36	0.22	0.63	0.82	0.75	0.66	0.91	0.07	0.17	0.15	-0.08	-0.26	-0.06
C 37	0.11	0.47	0.75	0.85	0.95	1.00	-0.06	-0.02	0.04	0.00	0.02	0.03
D 38	0.45	0.88	0.97	0.99	0.98	1.00	-0.07	0.02	0.04	0.02	-0.01	0.00
D 39	0.23	0.66	0.77	0.93	0.94	0.97	-0.02	0.05	-0.02	0.03	-0.02	-0.01
D 40	0.47	0.84	0.81	0.89	0.87	0.98	0.12	0.11	-0.06	-0.05	-0.10	-0.01
D 41	0.61	0.86	0.95	0.98	0.99	1.00	0.03	-0.03	-0.00	0.00	-0.00	0.00
D 42	0.16	0.51	0.72	0.83	0.86	0.93	0.01	0.05	0.05	-0.01	-0.06	-0.04
D 43	0.26	0.64	0.83	0.90	0.97	0.97	-0.00	0.00	0.02	-0.01	0.01	-0.01
D 44	0.49	0.78	0.83	0.85	0.87	0.90	0.18	0.09	-0.01	-0.07	-0.10	-0.09
D 45	0.29	0.78	0.85	0.93	0.99	0.97	-0.04	0.06	-0.00	-0.01	0.02	-0.02
D 46	0.06	0.30	0.70	0.95	0.96	0.99	-0.09	-0.15	0.04	0.13	0.05	0.03
D 47	0.50	0.85	0.91	0.93	0.97	0.98	0.05	0.03	-0.01	-0.03	-0.01	-0.01
D 48	0.40	0.57	0.67	0.82	0.89	0.99	0.19	0.01	-0.08	-0.07	-0.05	0.01
D 49	0.23	0.36	0.59	0.80	0.93	1.00	0.09	-0.07	-0.05	-0.02	0.02	0.04
D 50	0.04	0.12	0.32	0.80	0.95	1.00	-0.03	-0.14	-0.14	0.12	0.13	0.08

	ITEM CHARACTERISTIC CURVE						DEPARTURE FROM EXPECTED ICC					
ITEM NAME	SUBGROUP 1	2	3	4	5	6	SUBGROUP 1	2	3	4	5	6
D 51	0.37	0.80	0.97	1.00	0.99	1.00	-0.08	-0.02	0.06	0.04	0.01	0.01
D 52	0.57	0.78	0.92	0.97	0.98	1.00	0.08	-0.07	-0.01	0.00	-0.01	0.01
E 53	0.13	0.53	0.83	0.85	0.93	0.96	-0.06	-0.00	0.10	-0.01	-0.01	-0.01
E 54	0.08	0.26	0.30	0.42	0.62	0.70	0.05	0.13	0.04	-0.05	-0.04	-0.13
E 55	0.49	0.63	0.67	0.96	0.99	0.99	0.18	-0.06	-0.17	0.03	0.02	0.00
E 56	0.20	0.82	0.93	0.99	0.99	0.99	-0.17	0.07	0.05	0.04	0.01	-0.00
E 57	0.31	0.87	0.91	0.96	0.98	0.99	-0.10	0.09	0.02	0.01	0.00	-0.00
E 58	0.30	0.80	0.96	0.96	0.97	0.99	-0.09	0.03	0.07	0.01	-0.01	-0.00
E 59	0.01	0.06	0.16	0.29	0.47	0.84	-0.01	-0.02	-0.01	-0.03	-0.04	0.11
E 60	0.47	0.81	0.90	0.95	0.97	0.99	0.03	0.00	-0.01	-0.01	-0.01	-0.00
E 61	0.08	0.38	0.53	0.76	0.88	0.97	-0.02	0.03	-0.03	-0.00	0.00	0.02
E 62	0.05	0.10	0.29	0.57	0.84	0.92	0.00	-0.08	-0.06	0.01	0.09	0.04
E 63	0.18	0.37	0.59	0.85	0.94	0.99	0.04	-0.06	-0.06	0.03	0.03	0.03
E 64	0.00	0.09	0.35	0.58	0.64	0.79	-0.03	-0.05	0.07	0.10	-0.04	-0.05
F 65	0.44	0.73	0.72	0.70	0.87	0.96	0.21	0.14	-0.05	-0.19	-0.08	-0.02
F 66	0.06	0.15	0.50	0.61	0.79	0.96	-0.00	-0.08	0.07	-0.03	-0.01	0.05
F 67	0.25	0.86	0.86	0.82	0.73	0.82	0.04	0.30	0.11	-0.07	-0.22	-0.15
F 68	0.14	0.26	0.48	0.66	0.82	0.96	0.06	-0.02	-0.01	-0.04	-0.02	0.03
F 69	0.02	0.24	0.48	0.77	0.78	0.98	-0.06	-0.03	0.00	0.08	-0.05	0.05
F 70	0.28	0.48	0.68	0.80	0.91	1.00	0.10	-0.02	-0.03	-0.06	-0.02	0.03
F 71	0.29	0.77	0.83	0.95	0.96	1.00	-0.04	0.06	-0.02	0.01	-0.01	0.01
F 72	0.24	0.63	0.84	0.85	0.95	0.98	-0.01	0.01	0.05	-0.05	-0.00	-0.00
F 73	0.57	0.85	0.85	0.83	0.85	0.97	0.21	0.10	-0.03	-0.12	-0.13	-0.02
F 74	0.09	0.36	0.55	0.73	0.96	0.98	-0.02	-0.00	-0.03	-0.04	0.08	0.03
F 75	0.02	0.40	0.66	0.84	0.99	1.00	-0.12	-0.03	0.01	0.03	0.08	0.04
F 76	0.04	0.13	0.25	0.54	0.82	0.95	-0.01	-0.05	-0.09	-0.01	0.08	0.07
G 77	0.05	0.18	0.24	0.58	0.85	0.96	-0.00	-0.02	-0.13	-0.01	0.09	0.07
G 78	0.01	0.06	0.10	0.34	0.74	0.96	-0.02	-0.05	-0.13	-0.08	0.12	0.16
G 79	0.01	0.14	0.31	0.46	0.78	0.97	-0.03	-0.03	-0.01	-0.08	0.06	0.10
G 80	0.38	0.66	0.87	0.95	0.98	1.00	0.04	-0.07	0.01	0.01	0.01	0.01
G 81	0.48	0.85	0.81	0.85	0.97	0.99	0.10	0.09	-0.08	-0.09	-0.01	-0.00
G 82	0.04	0.09	0.26	0.39	0.41	0.61	0.02	0.02	0.10	0.07	-0.10	-0.12
G 83	0.21	0.84	0.95	0.98	0.98	0.99	-0.17	0.08	0.07	0.03	0.00	-0.00
G 84	0.49	0.85	0.94	0.98	0.99	0.99	-0.01	0.00	0.01	0.01	0.00	-0.01
G 85	0.39	0.80	0.91	0.99	1.00	1.00	-0.05	-0.01	0.00	0.03	0.02	0.01
G 86	0.00	0.00	0.15	0.38	0.80	0.96	-0.03	-0.12	-0.10	-0.06	0.16	0.15
G 87	0.15	0.36	0.76	0.91	0.90	0.97	-0.01	-0.11	0.07	0.07	-0.02	0.00
H 88	0.35	0.75	0.82	0.92	0.98	0.98	0.02	0.04	-0.04	-0.01	0.01	-0.01
H 89	0.00	0.07	0.16	0.14	0.18	0.52	-0.01	0.04	0.09	-0.02	-0.11	0.00
H 90	0.13	0.59	0.63	0.75	0.93	0.95	-0.02	0.14	-0.03	-0.08	0.01	-0.02
H 91	0.01	0.07	0.25	0.60	0.83	0.94	-0.03	-0.10	-0.08	0.05	0.10	0.07
H 92	0.03	0.05	0.16	0.18	0.32	0.49	0.02	0.01	0.07	-0.01	-0.02	-0.08
H 93	0.27	0.79	0.89	0.98	1.00	0.99	-0.10	0.04	0.01	0.03	0.02	-0.00
H 94	0.06	0.23	0.37	0.50	0.63	0.65	0.03	0.09	0.09	0.02	-0.05	-0.19
H 95	0.01	0.05	0.16	0.26	0.39	0.45	0.00	0.01	0.06	0.05	0.02	-0.15
H 96	0.05	0.42	0.74	0.88	0.96	0.97	-0.11	-0.04	0.06	0.05	0.04	0.00
H 97	0.08	0.39	0.73	0.70	0.84	0.91	-0.03	0.03	0.15	-0.07	-0.04	-0.04
I 98	0.01	0.01	0.05	0.05	0.15	0.22	0.01	-0.00	0.02	-0.01	0.03	-0.06
I 99	0.11	0.68	0.91	0.87	0.87	0.96	-0.12	0.09	0.14	-0.02	-0.08	-0.02
I100	0.00	0.03	0.27	0.74	0.88	0.95	-0.05	-0.17	-0.10	0.15	0.12	0.06

ITEM NAME	ITEM CHARACTERISTIC CURVE SUBGROUP 1	2	3	4	5	6	DEPARTURE FROM EXPECTED ICC SUBGROUP 1	2	3	4	5	6
I101	0.04	0.04	0.07	0.11	0.36	0.49	0.03	0.01	-0.01	-0.06	0.06	-0.04
I102	0.12	0.52	0.78	0.98	0.97	1.00	-0.10	-0.05	0.02	0.09	0.02	0.02
I103	0.03	0.34	0.68	0.86	0.97	0.93	-0.10	-0.06	0.06	0.06	0.07	-0.03
I104	0.01	0.02	0.17	0.44	0.57	0.70	-0.01	-0.06	-0.01	0.09	0.03	-0.05
I105	0.01	0.12	0.26	0.65	0.89	0.96	-0.04	-0.08	-0.11	0.06	0.12	0.07
I106	0.01	0.21	0.52	0.73	0.83	0.91	-0.06	-0.05	0.07	0.05	0.01	-0.01
I107	0.17	0.48	0.88	0.97	1.00	0.99	-0.08	-0.13	0.09	0.07	0.04	0.01
I108	0.12	0.48	0.83	0.93	0.96	0.98	-0.08	-0.07	0.08	0.05	0.02	0.00
J109	0.00	0.00	0.02	0.08	0.09	0.34	-0.00	-0.01	-0.01	0.01	-0.04	0.05
J110	0.02	0.02	0.10	0.35	0.65	0.73	0.00	-0.06	-0.07	0.02	0.12	-0.01
J111	0.00	0.04	0.02	0.00	0.03	0.23	-0.00	0.03	0.00	-0.04	-0.05	0.04
J112	0.04	0.06	0.17	0.34	0.44	0.62	0.02	-0.00	0.02	0.05	-0.03	-0.08
J113	0.24	0.63	0.79	0.80	0.89	0.99	0.03	0.07	0.04	-0.09	-0.05	0.01
J114	0.35	0.88	0.99	1.00	0.97	1.00	-0.13	0.05	0.07	0.03	-0.02	0.01
J115	0.02	0.09	0.19	0.32	0.41	0.68	0.00	0.02	0.05	0.02	-0.07	-0.03
J116	0.03	0.19	0.65	0.97	1.00	0.99	-0.10	-0.22	0.02	0.17	0.10	0.03
J117	0.14	0.32	0.50	0.78	0.76	0.87	0.06	0.03	0.00	0.07	-0.09	-0.07
J118	0.00	0.09	0.29	0.33	0.54	0.89	-0.02	-0.01	0.08	-0.07	-0.06	0.10
J119	0.07	0.37	0.53	0.83	0.89	0.94	-0.04	0.02	-0.04	0.06	0.01	-0.01
K120	0.24	0.61	0.71	0.82	0.91	0.99	0.04	0.07	-0.03	-0.06	-0.03	0.01
K121	0.14	0.24	0.48	0.55	0.76	0.96	0.07	-0.00	0.04	-0.10	-0.05	0.05
K122	0.00	0.00	0.03	0.16	0.46	0.70	-0.01	-0.04	-0.07	-0.06	0.09	0.09
K123	0.05	0.20	0.28	0.40	0.49	0.75	0.02	0.09	0.06	-0.02	-0.12	-0.05
K124	0.00	0.00	0.06	0.34	0.66	0.87	-0.02	-0.08	-0.12	-0.01	0.12	0.12
K125	0.04	0.32	0.46	0.63	0.78	0.87	-0.02	0.08	0.03	-0.02	-0.03	-0.05
K126	0.02	0.12	0.26	0.44	0.69	0.86	-0.01	-0.01	-0.00	-0.03	0.03	0.03
K127	0.07	0.43	0.74	0.95	0.97	0.99	-0.11	-0.07	0.03	0.09	0.04	0.02
K128	0.00	0.03	0.09	0.34	0.60	0.78	-0.02	-0.05	-0.08	0.01	0.08	0.05
K129	0.05	0.25	0.58	0.61	0.87	0.89	-0.03	-0.02	0.11	-0.07	0.04	-0.04
L130	0.26	0.65	0.82	0.94	0.98	0.97	-0.02	-0.00	-0.00	0.02	0.02	-0.02
L131	0.00	0.00	0.06	0.34	0.69	0.84	-0.02	-0.08	-0.12	-0.01	0.15	0.09
L132	0.03	0.37	0.74	0.96	0.99	1.00	-0.14	-0.11	0.04	0.11	0.06	0.03
L133	0.00	0.01	0.02	0.03	0.11	0.55	-0.00	-0.01	-0.02	-0.06	-0.08	0.17
L134	0.00	0.01	0.02	0.19	0.58	0.82	-0.01	-0.05	-0.11	-0.08	0.12	0.14
L135	0.02	0.35	0.58	0.90	0.94	0.98	-0.10	-0.04	-0.03	0.11	0.04	0.02
L136	0.00	0.07	0.20	0.27	0.33	0.24	-0.01	0.04	0.13	0.10	0.02	-0.30
L137	0.03	0.17	0.54	0.90	0.96	0.98	-0.07	-0.18	-0.02	0.15	0.09	0.03
L138	0.00	0.01	0.04	0.21	0.25	0.61	-0.01	-0.02	-0.04	0.05	-0.06	0.07
L139	0.00	0.00	0.02	0.11	0.27	0.32	-0.00	-0.02	-0.02	0.01	0.08	-0.06
L140	0.15	0.71	0.84	0.87	0.90	0.95	-0.08	0.12	0.07	-0.02	-0.05	-0.03
L141	0.07	0.46	0.62	0.58	0.76	0.78	-0.01	0.19	0.14	-0.11	-0.08	-0.14

GROUP	SCORE RANGE	MEAN ABILITY	NO. IN SUBGROUP
1	1 - 50	-2.12	101
2	51 - 75	-0.33	99
3	76 - 93	0.56	103
4	94 - 108	1.46	103
5	109 - 120	2.28	99
6	121 - 140	3.25	97

N = 602

E.4 Cloze-Type Test (Malaysian Group): Item Fit Statistics

(Items ordered by total fit-t; 6 misfitting persons omitted)

ITEM NAME	ITEM DIFFIC.	BETWEEN GROUP FIT-T	TOTAL FIT-T	RASCH DISCRIM. INDEX
J116	0.04	6.43	-8.10	1.50
G 86	1.72	5.57	-7.54	1.43
I100	1.09	5.56	-7.13	1.47
L137	0.32	5.03	-7.11	1.43
D 50	0.72	5.46	-6.83	1.45
B 15	-0.89	5.50	-6.79	1.38
L132	-0.27	4.97	-6.43	1.41
G 78	1.81	4.78	-6.35	1.35
L134	2.46	4.63	-5.81	1.30
I105	1.07	3.87	-5.54	1.37
K124	2.11	4.44	-5.39	1.33
L131	2.11	4.43	-5.27	1.31
D 46	-0.11	4.25	-5.27	1.35
B 20	0.80	5.12	-5.02	1.33
H 91	1.27	3.36	-5.01	1.34
F 75	-0.06	3.63	-4.90	1.34
C 29	-0.43	3.69	-4.87	1.29
K127	-0.34	3.25	-4.87	1.29
L135	0.11	3.31	-4.69	1.32
C 27	0.83	2.86	-4.55	1.34
I107	-0.80	3.72	-4.30	1.24
E 56	-1.45	3.49	-4.11	1.21
I102	-0.61	2.98	-4.04	1.25
H 96	-0.19	2.41	-3.93	1.24
B 23	-2.11	2.69	-3.85	1.17
B 12	-1.78	3.33	-3.82	1.20
G 83	-1.50	3.51	-3.71	1.19
I103	0.06	3.31	-3.60	1.24
I108	-0.53	2.07	-3.49	1.19
E 62	1.20	1.83	-3.39	1.20
F 76	1.24	2.04	-3.35	1.22
B 21	-2.03	2.03	-3.22	1.15
C 26	-1.80	2.22	-3.18	1.16
K122	2.79	3.04	-3.10	1.21
G 77	1.10	2.51	-3.06	1.21
J114	-1.96	3.20	-2.94	1.15
H 93	-1.45	1.58	-2.88	1.14
G 79	1.31	2.51	-2.86	1.23
K128	2.20	2.03	-2.77	1.19
A 5	1.20	1.89	-2.68	1.17
L133	3.77	3.36	-2.67	1.16
A 4	1.78	1.73	-2.65	1.14
J110	2.17	2.22	-2.50	1.13
D 38	-2.14	0.77	-2.35	1.08
C 37	-0.32	0.99	-2.26	1.16
C 30	-1.76	1.11	-2.10	1.07
E 57	-1.63	1.34	-2.06	1.09

(Items ordered by total fit-t; 6 misfitting persons omitted)

ITEM NAME	ITEM DIFFIC.	BETWEEN GROUP FIT-T	TOTAL FIT-T	RASCH DISCRIM. INDEX
I106	0.73	1.60	-2.01	1.14
F 69	0.67	2.23	-2.00	1.19
E 58	-1.56	1.55	-1.99	1.10
E 59	2.21	0.90	-1.93	1.11
A 2	-3.33	0.46	-1.80	1.09
D 51	-1.84	1.90	-1.75	1.14
G 87	-0.22	1.90	-1.53	1.09
G 85	-1.76	0.50	-1.48	1.09
K126	1.60	-1.26	-1.43	1.07
J118	1.87	2.42	-1.43	1.08
F 74	0.23	1.31	-1.40	1.13
C 28	2.64	0.36	-1.30	1.06
L138	3.09	1.39	-1.30	1.09
C 33	-3.46	-1.09	-1.27	1.04
I 99	-0.67	4.91	-1.21	1.01
F 66	0.88	1.37	-1.12	1.10
E 53	-0.43	1.17	-1.08	1.05
A 10	-3.37	-0.01	-1.08	1.07
F 71	-1.25	0.15	-1.05	1.04
D 45	-1.27	0.87	-0.98	1.02
E 63	-0.07	1.06	-0.96	1.06
J119	0.27	0.10	-0.92	1.08
G 84	-2.07	-1.63	-0.78	1.02
L130	-0.96	-0.39	-0.74	1.03
A 3	-2.59	-1.09	-0.62	1.03
A 9	-3.10	3.16	-0.54	0.98
B 19	-1.72	0.42	-0.52	1.03
J109	4.16	0.12	-0.39	1.06
I104	2.12	1.67	-0.38	1.05
L140	-0.69	3.39	-0.35	0.96
J111	4.79	3.79	-0.26	0.97
E 61	0.31	-0.75	-0.24	1.06
D 43	-0.89	-1.18	-0.15	0.99
F 72	-0.79	0.15	-0.06	0.97
D 39	-0.79	-0.24	-0.01	0.99
C 25	-3.86	2.79	0.11	0.91
K129	0.68	1.83	0.16	0.99
L139	3.75	1.37	0.17	1.01
E 60	-1.76	-1.15	0.33	0.94
I101	3.11	2.92	0.37	0.92
I 98	4.26	0.90	0.38	0.91
H 88	-1.23	-0.33	0.44	0.95
G 80	-1.32	0.08	0.46	1.01
D 49	-0.04	2.28	0.58	0.96
E 64	1.52	2.26	0.59	1.02
B 22	-2.18	-1.59	0.61	0.99
A 8	-2.62	0.20	0.66	0.93
F 68	0.62	0.94	0.71	0.92
A 11	-2.59	0.04	0.93	1.01
D 41	-2.45	-1.06	0.98	0.97
H 89	3.18	3.48	1.03	0.89
H 90	-0.11	2.13	1.09	0.91

(Items ordered by total fit-t; 6 misfitting persons omitted)

ITEM NAME	ITEM DIFFIC.	BETWEEN GROUP FIT-T	TOTAL FIT-T	RASCH DISCRIM. INDEX
B 18	-2.38	0.94	1.23	0.92
D 47	-1.88	1.32	1.25	0.89
A 1	-3.41	3.31	1.43	0.83
D 42	-0.15	1.59	1.44	0.85
D 52	-2.05	0.87	1.47	0.93
H 97	0.25	2.93	1.63	0.89
K125	0.85	1.07	1.64	0.89
C 31	-2.73	3.75	1.75	0.81
H 92	2.93	2.21	1.75	0.84
J113	-0.56	2.55	1.77	0.86
C 32	-1.84	3.42	1.79	0.81
J115	2.33	0.35	1.80	0.89
J112	2.39	1.31	1.87	0.87
B 16	0.85	2.90	1.89	0.79
K121	0.82	2.79	1.95	0.84
B 14	-0.11	3.12	1.97	0.91
K120	-0.48	1.24	2.23	0.88
H 95	2.81	2.36	2.28	0.84
B 24	-1.03	1.83	2.29	0.86
C 35	-0.09	2.17	2.32	0.80
F 70	-0.33	2.29	2.41	0.85
G 81	-1.48	4.53	2.71	0.79
J117	0.56	2.93	2.74	0.74
E 55	-1.12	5.35	2.90	0.83
G 82	2.22	3.51	3.60	0.75
D 40	-1.33	6.51	3.77	0.69
K123	1.82	3.06	3.91	0.74
A 6	-1.03	4.49	4.50	0.70
D 48	-0.56	4.79	4.81	0.67
C 34	-1.40	15.07	4.96	0.38
B 17	0.88	4.58	5.18	0.57
L136	3.08	7.12	5.32	0.66
E 54	1.60	4.73	5.38	0.60
H 94	1.53	5.09	5.39	0.59
F 73	-1.45	9.55	5.46	0.53
F 67	-0.59	12.70	5.67	0.39
D 44	-1.12	9.11	5.82	0.51
L141	0.65	6.63	5.89	0.55
C 36	-0.13	9.43	6.35	0.44
A 7	0.54	10.99	6.66	0.43
F 65	-0.67	8.06	7.60	0.46
B 13	-1.03	15.05	11.71	-0.02

TOTAL FIT-T:
Mean = -0.57
SD = 3.51
Range = -8.10 to 11.71

BETWEEN-GROUP FIT-T:
Mean = 2.84
SD = 2.83
Range = -1.63 to 15.07

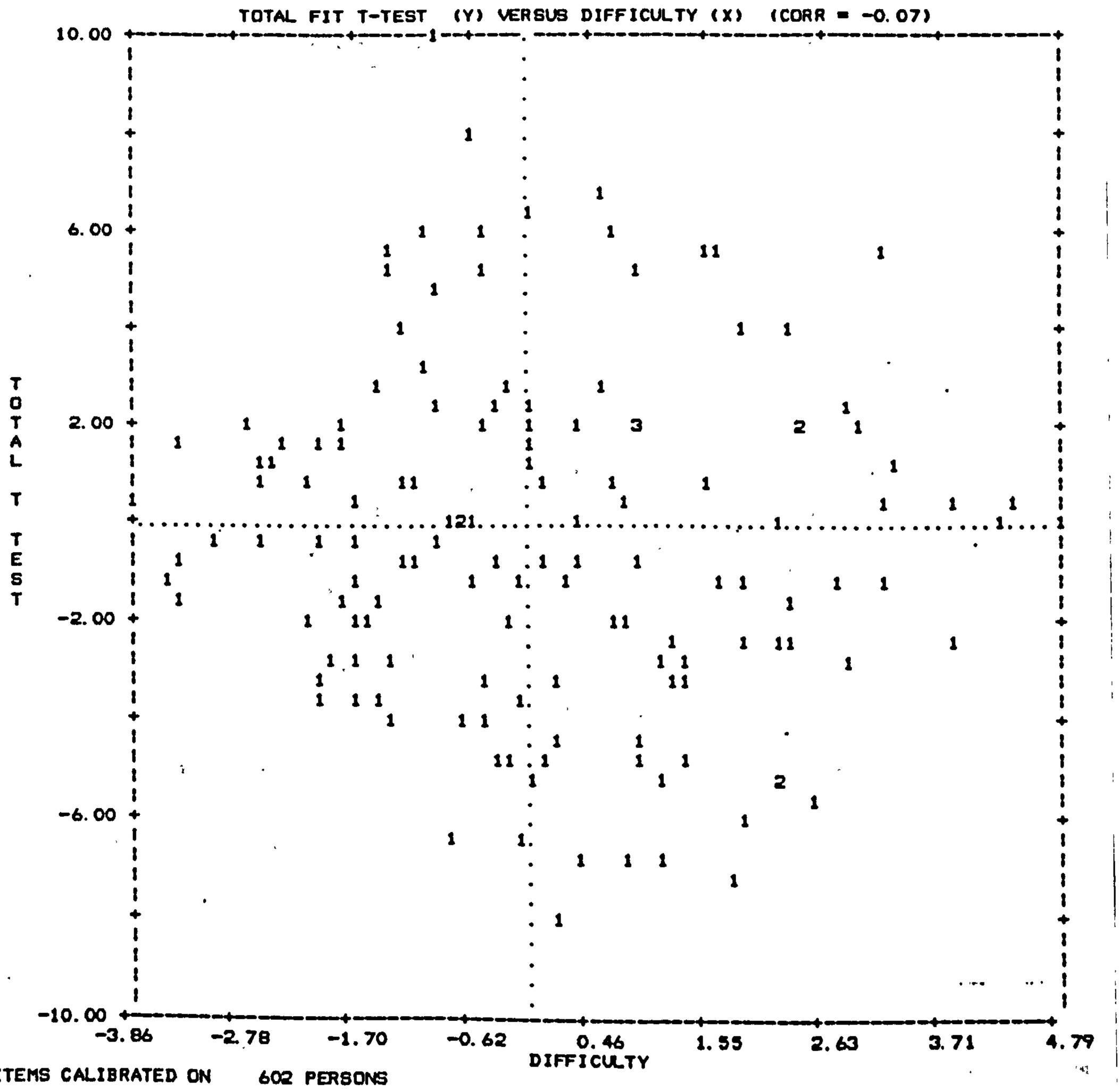
DISCRIM. INDEX:
Range = -0.12 to 1.50

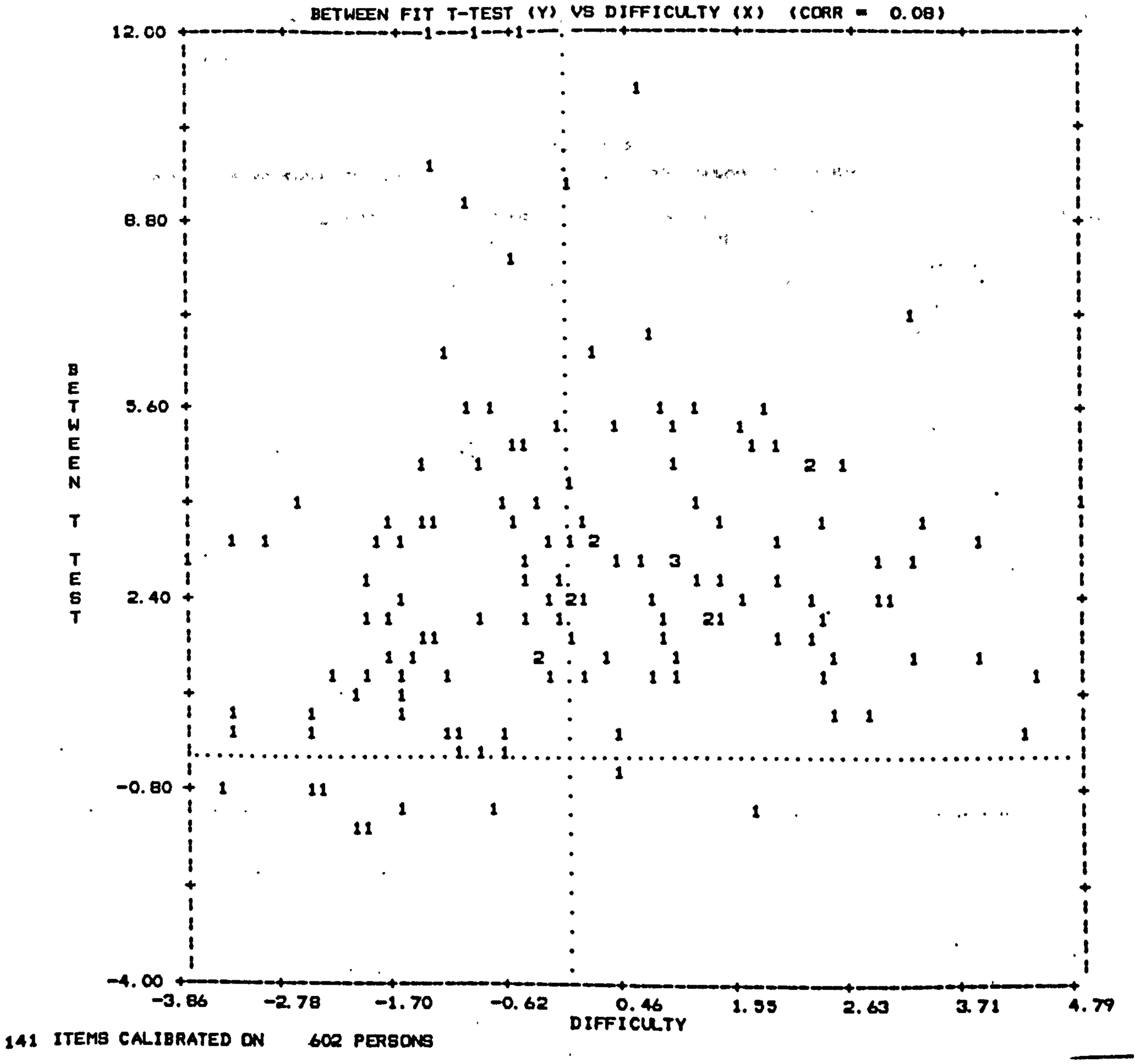
E.5 Person Statistics and Standardized Residuals for Misfitting Persons (Malaysian Group)

Person No.	Raw Score	Abil. Estim.	Total Fit-t	Standardized Residuals (letters refer to passages)
52	59	-0.58	2.98	**A :** -3 -3 -2 0 0 -1 0 -2 -3 0 0 **B :** -1 -1 1 -1 0 0 -2 0 1 -2 0 0 -1 **C :** 0 -1 0 0 0 0 0 -1 0 0 1 1 1 **D :** 0 -1 0 -2 1 0 0 0 0 0 0 1 0 0 0 **E :** 0 0 0 0 0 -1 3 0 0 0 1 0 **F :** -1 2 0 0 0 0 0 -1 0 0 0 0 **G :** 0 0 0 0 0 4 0 0 0 0 1 **H :** 0 0 1 0 0 0 0 0 0 0 **I :** 0 0 0 0 0 1 3 2 1 0 0 **J :** 0 0 0 0 0 0 4 1 0 0 0 **K :** 1 0 0 0 0 0 0 1 3 0 **L :** 0 0 0 0 0 0 0 0 0 0 0 0
505	97	1.18	2.92	**A :** 0 0 0 0 0 0 0 0 0 0 0 **B :** 0 0 0 0 0 -1 0 -4 0 0 -5 -5 0 **C :** 0 -4 0 0 0 0 0 0 0 0 0 0 -2 **D :** 0 0 0 0 0 -2 -3 0 0 0 0 0 0 0 0 **E :** 0 1 0 0 0 0 1 0 0 1 -1 0 **F :** 0 -1 0 -1 0 -2 -3 -2 0 -1 0 1 **G :** 0 1 1 0 -3 0 -3 0 0 0 0 **H :** 0 0 0 1 2 0 1 2 0 -1 **I :** 0 0 0 0 0 0 1 0 -1 0 0 **J :** 0 0 0 1 0 0 1 0 0 1 0 **K :** 0 -1 0 0 0 -1 1 0 1 -1 **L :** 0 0 0 0 0 -1 0 0 0 0 0 -1
66	83	0.50	2.68	**A :** 0 0 0 0 1 -2 0 0 0 0 -4 **B :** -3 0 0 0 1 0 -4 0 0 0 0 0 0 **C :** 0 0 0 0 0 0 0 -3 0 -2 -1 0 -1 **D :** 0 0 0 0 -1 0 0 0 -1 0 0 0 0 0 0 **E :** -1 1 0 0 0 -2 0 -3 -1 0 0 1 **F :** 0 1 0 0 0 0 0 -1 0 0 0 1 **G :** 1 1 1 0 0 2 0 0 0 0 -1 **H :** 0 0 0 0 3 0 0 0 0 0 **I :** 0 0 0 0 -1 0 2 1 0 0 0 **J :** 0 0 0 0 0 0 0 0 0 0 0 **K :** -1 0 0 1 0 0 1 0 2 1 **L :** 0 0 -1 0 0 -1 0 0 0 0 -1 1

Person No.	Raw Score	Abil. Estim.	Total Fit-t	Standardized Residuals (letters refer to passages)
564	28	-2.12	2.36	A : 0 0 0 0 0 0 0 -1 -1 -1 0 B : 1 0 0 0 0 0 -1 0 0 0 -1 0 1 C : -2 0 0 0 0 0 -1 0 -1 0 0 0 0 D : 0 0 0 0 0 0 0 1 0 0 2 2 0 1 1 E : 0 6 0 1 0 0 0 1 0 0 0 0 F : 2 0 0 0 0 0 0 0 1 3 2 0 G : 0 0 0 0 0 0 0 0 0 0 0 0 H : 0 0 0 0 0 1 0 0 0 0 I : 0 2 0 0 2 0 0 0 0 1 2 J : 0 0 0 9 2 1 0 0 0 0 3 K : 0 0 0 0 0 0 0 0 0 0 0 L : 0 0 0 0 0 0 0 0 0 0 0 0 0
436	24	-2.38	2.25	A : -1 -1 -1 0 0 0 4 0 -1 0 0 B : 0 0 3 0 0 5 1 0 0 0 0 0 0 C : -2 0 0 0 0 0 0 1 0 0 3 0 0 D : 0 0 1 -1 0 2 1 0 0 0 0 0 0 0 0 E : 0 0 0 0 0 0 0 0 3 0 0 0 F : 2 0 0 0 0 2 0 0 1 0 0 0 G : 0 0 0 1 1 0 0 0 0 0 0 H : 0 0 0 0 0 0 0 0 0 3 I : 0 0 0 0 0 0 0 0 0 0 0 J : 0 0 0 0 0 0 0 0 0 0 3 K : 2 4 0 0 0 0 0 0 0 0 L : 0 0 0 0 0 0 0 0 0 0 0 0
4	102	1.45	2.17	A : 0 0 -7 1 0 0 0 0 0 0 0 B : 0 -3 0 0 -1 0 0 0 0 0 0 0 0 C : 0 0 0 0 0 0 0 -5 0 0 0 0 -2 D : -5 -3 0 0 0 0 -3 0 0 0 0 0 0 0 0 E : 0 0 0 0 0 0 1 0 0 0 0 1 F : -2 -1 -2 -1 0 0 0 0 0 0 0 -1 G : 0 1 0 0 -4 0 0 0 0 1 0 H : 0 0 -2 0 0 0 0 0 0 -1 I : 0 0 0 0 0 0 0 0 0 0 0 -2 J : 0 1 0 0 0 0 1 0 0 1 -1 K : 0 -1 0 0 1 0 0 0 1 0 L : 0 0 0 0 1 0 0 0 2 0 0 0

E.6 Cloze-Type Test (Malaysian Group): Item Fit Statistics vs Item Difficulty





141 ITEMS CALIBRATED ON 602 PERSONS

APPENDIX F
RASCH STATISTICS FOR CLOZE-TYPE TEST (TANZANIAN DATA)

F.1 Cloze-Type Test (Tanzanian Group): Raw Scores, Rasch Ability Estimates and Standard Errors

TABLE 1
(All measurable persons included)

RAW SCORE	FREQ. COUNT	ABILITY ESTIM.	STANDARD ERROR
140	0	7.03	1.09
139	0	6.20	0.81
138	0	5.66	0.68
137	0	5.26	0.60
136	0	4.94	0.55
135	0	4.67	0.50
134	0	4.44	0.47
133	0	4.24	0.44
132	0	4.05	0.42
131	0	3.89	0.40
130	0	3.74	0.38
129	1	3.60	0.37
128	0	3.47	0.36
127	0	3.35	0.35
126	0	3.24	0.34
125	0	3.13	0.33
124	0	3.03	0.32
123	0	2.93	0.31
122	0	2.84	0.30
121	0	2.75	0.30
120	0	2.66	0.29
119	1	2.58	0.29
118	1	2.50	0.28
117	0	2.42	0.28
116	1	2.35	0.27
115	0	2.27	0.27
114	0	2.20	0.27
113	0	2.14	0.26
112	2	2.07	0.26
111	1	2.00	0.26
110	3	1.94	0.25
109	3	1.87	0.25
108	2	1.81	0.25
107	3	1.75	0.25
106	2	1.69	0.24
105	2	1.63	0.24
104	1	1.58	0.24
103	2	1.52	0.24
102	3	1.46	0.24
101	1	1.41	0.24
100	3	1.35	0.23
99	3	1.30	0.23
98	2	1.24	0.23
97	4	1.19	0.23
96	3	1.14	0.23
95	0	1.09	0.23
94	3	1.04	0.23
93	1	0.99	0.23
92	1	0.93	0.23
91	4	0.88	0.22

TABLE 2
(7 misfitting persons excluded)

RAW SCORE	FREQ. COUNT	ABILITY ESTIM.	STANDARD ERROR
140	0	7.09	1.09
139	0	6.26	0.81
138	0	5.72	0.68
137	0	5.32	0.61
136	0	4.99	0.55
135	0	4.72	0.51
134	0	4.49	0.47
133	0	4.28	0.45
132	0	4.09	0.42
131	0	3.93	0.40
130	0	3.78	0.39
129	1	3.63	0.37
128	0	3.50	0.36
127	0	3.38	0.35
126	0	3.27	0.34
125	0	3.16	0.33
124	0	3.06	0.32
123	0	2.96	0.31
122	0	2.86	0.31
121	0	2.77	0.30
120	0	2.69	0.29
119	1	2.61	0.29
118	1	2.53	0.28
117	0	2.45	0.28
116	1	2.37	0.27
115	0	2.30	0.27
114	0	2.23	0.27
113	0	2.16	0.26
112	2	2.09	0.26
111	1	2.02	0.26
110	3	1.96	0.26
109	3	1.90	0.25
108	2	1.83	0.25
107	3	1.77	0.25
106	2	1.71	0.25
105	2	1.65	0.24
104	1	1.59	0.24
103	2	1.54	0.24
102	3	1.48	0.24
101	1	1.42	0.24
100	3	1.37	0.24
99	3	1.31	0.23
98	2	1.26	0.23
97	4	1.21	0.23
96	3	1.15	0.23
95	0	1.10	0.23
94	3	1.05	0.23
93	1	1.00	0.23
92	1	0.95	0.23
91	3	0.89	0.23

TABLE 1
(All measurable persons included)

RAW SCORE	FREQ. COUNT	ABILITY ESTIM.	STANDARD ERROR
90	1	0.83	0.22
89	1	0.79	0.22
88	2	0.74	0.22
87	2	0.69	0.22
86	2	0.64	0.22
85	2	0.59	0.22
84	3	0.54	0.22
83	1	0.49	0.22
82	3	0.45	0.22
81	2	0.40	0.22
80	4	0.35	0.22
79	0	0.31	0.22
78	4	0.26	0.22
77	4	0.21	0.22
76	2	0.17	0.22
75	3	0.12	0.22
74	5	0.07	0.22
73	1	0.03	0.22
72	0	-0.02	0.22
71	4	-0.06	0.21
70	0	-0.11	0.21
69	5	-0.16	0.21
68	2	-0.20	0.21
67	2	-0.25	0.21
66	1	-0.29	0.21
65	2	-0.34	0.21
64	3	-0.38	0.21
63	2	-0.43	0.21
62	3	-0.48	0.21
61	2	-0.52	0.21
60	1	-0.57	0.21
59	1	-0.61	0.22
58	3	-0.66	0.22
57	4	-0.71	0.22
56	3	-0.75	0.22
55	0	-0.80	0.22
54	3	-0.84	0.22
53	3	-0.89	0.22
52	1	-0.94	0.22
51	2	-0.98	0.22
50	1	-1.03	0.22
49	5	-1.08	0.22
48	0	-1.13	0.22
47	2	-1.17	0.22
46	3	-1.22	0.22
45	4	-1.27	0.22
44	1	-1.32	0.22
43	3	-1.37	0.22
42	1	-1.42	0.22
41	1	-1.47	0.23

TABLE 2
(7 misfitting persons excluded)

RAW SCORE	FREQ. COUNT	ABILITY ESTIM.	STANDARD ERROR
90	1	0.84	0.23
89	1	0.79	0.22
88	2	0.74	0.22
87	2	0.70	0.22
86	2	0.65	0.22
85	2	0.60	0.22
84	3	0.55	0.22
83	1	0.50	0.22
82	2	0.45	0.22
81	2	0.40	0.22
80	4	0.36	0.22
79	0	0.31	0.22
78	3	0.26	0.22
77	4	0.22	0.22
76	2	0.17	0.22
75	3	0.12	0.22
74	4	0.07	0.22
73	1	0.03	0.22
72	0	-0.02	0.22
71	4	-0.07	0.22
70	0	-0.11	0.22
69	5	-0.16	0.22
68	2	-0.20	0.22
67	2	-0.25	0.22
66	1	-0.30	0.22
65	2	-0.34	0.22
64	3	-0.39	0.22
63	2	-0.44	0.22
62	2	-0.48	0.22
61	2	-0.53	0.22
60	1	-0.58	0.22
59	1	-0.62	0.22
58	3	-0.67	0.22
57	4	-0.71	0.22
56	3	-0.76	0.22
55	0	-0.81	0.22
54	3	-0.86	0.22
53	3	-0.90	0.22
52	1	-0.95	0.22
51	2	-1.00	0.22
50	1	-1.05	0.22
49	5	-1.09	0.22
48	0	-1.14	0.22
47	2	-1.19	0.22
46	3	-1.24	0.22
45	4	-1.29	0.22
44	1	-1.34	0.22
43	3	-1.39	0.22
42	1	-1.44	0.23
41	1	-1.49	0.23

TABLE 1
(All measurable persons included)

RAW SCORE	FREQ. COUNT	ABILITY ESTIM.	STANDARD ERROR
40	2	-1.52	0.23
39	1	-1.57	0.23
38	2	-1.62	0.23
37	4	-1.68	0.23
36	1	-1.73	0.23
35	3	-1.78	0.23
34	3	-1.84	0.24
33	3	-1.89	0.24
32	0	-1.95	0.24
31	4	-2.01	0.24
30	0	-2.06	0.24
29	1	-2.12	0.25
28	0	-2.18	0.25
27	2	-2.25	0.25
26	4	-2.31	0.25
25	2	-2.37	0.26
24	3	-2.44	0.26
23	4	-2.51	0.26
22	4	-2.58	0.27
21	1	-2.65	0.27
20	2	-2.72	0.28
19	1	-2.80	0.28
18	3	-2.88	0.29
17	2	-2.96	0.29
16	1	-3.05	0.30
15	3	-3.14	0.31
14	3	-3.23	0.32
13	0	-3.33	0.33
12	3	-3.44	0.34
11	5	-3.55	0.35
10	1	-3.68	0.36
9	0	-3.81	0.38
8	4	-3.95	0.40
7	1	-4.12	0.41
6	1	-4.30	0.44
5	0	-4.51	0.48
4	5	-4.76	0.53
3	1	-5.08	0.60
2	0	-5.51	0.73
1	0	-6.23	1.02

No. of persons = 243
Mean ability = -0.72
SD ability = 1.78
Group ability range: -5.08 to 3.60

Person separability index = 0.98
No. of person strata = 9.31

TABLE 2
(7 misfitting persons excluded)

RAW SCORE	FREQ. COUNT	ABILITY ESTIM.	STANDARD ERROR
40	1	-1.54	0.23
39	1	-1.59	0.23
38	1	-1.64	0.23
37	4	-1.70	0.23
36	1	-1.75	0.23
35	3	-1.80	0.23
34	3	-1.86	0.24
33	3	-1.92	0.24
32	0	-1.97	0.24
31	4	-2.03	0.24
30	0	-2.09	0.24
29	1	-2.15	0.25
28	0	-2.21	0.25
27	2	-2.27	0.25
26	4	-2.33	0.26
25	2	-2.40	0.26
24	3	-2.47	0.26
23	4	-2.53	0.27
22	4	-2.60	0.27
21	1	-2.68	0.27
20	2	-2.75	0.28
19	1	-2.83	0.28
18	3	-2.91	0.29
17	2	-2.99	0.30
16	1	-3.08	0.30
15	3	-3.17	0.31
14	3	-3.26	0.32
13	0	-3.37	0.33
12	3	-3.47	0.34
11	5	-3.59	0.35
10	1	-3.71	0.36
9	0	-3.84	0.38
8	4	-3.99	0.39
7	1	-4.15	0.42
6	1	-4.33	0.44
5	0	-4.55	0.48
4	5	-4.80	0.53
3	1	-5.12	0.60
2	0	-5.55	0.73
1	0	-6.27	1.02

No. of persons = 236
Mean ability = -0.74
SD ability = 1.78
Group ability range: -5.12 to 3.63

Person separability index = 0.98
No. of person strata = 9.42

F.2 Cloze-Type Test (Tanzanian Group): Item Difficulty Estimates & Standard Errors

SET 1 (All measurable persons included)			SET 2 (7 misfitting persons excluded)	
ITEM NAME	ITEM DIFFICULTY	STANDARD ERROR	ITEM DIFFICULTY	STANDARD ERROR
A 1	-4.08	0.25	-4.16	0.26
A 2	-2.57	0.19	-2.63	0.19
A 3	-2.13	0.18	-2.14	0.18
A 4	3.91	0.40	3.91	0.40
A 5	0.56	0.17	0.49	0.17
A 6	-1.12	0.17	-1.23	0.17
A 7	1.03	0.18	1.01	0.18
A 8	-2.20	0.18	-2.24	0.19
A 9	-2.99	0.20	-3.11	0.21
A 10	-3.33	0.22	-3.34	0.22
A 11	-3.52	0.22	-3.58	0.23
B 12	-1.50	0.17	-1.49	0.17
B 13	-2.47	0.19	-2.56	0.19
B 14	-0.98	0.16	-1.03	0.17
B 15	-0.59	0.16	-0.56	0.17
B 16	0.97	0.18	0.92	0.18
B 17	0.15	0.16	0.10	0.17
B 18	-2.68	0.19	-2.75	0.20
B 19	-1.79	0.17	-1.82	0.18
B 20	2.34	0.23	2.37	0.23
B 21	-1.14	0.17	-1.17	0.17
B 22	-2.26	0.18	-2.28	0.19
B 23	-1.88	0.18	-1.89	0.18
B 24	-1.28	0.17	-1.40	0.17
C 25	-3.73	0.23	-3.80	0.24
C 26	-1.45	0.17	-1.49	0.17
C 27	-0.22	0.16	-0.21	0.17
C 28	1.00	0.18	0.98	0.18
C 29	-0.88	0.16	-0.87	0.17
C 30	-1.91	0.18	-1.92	0.18
C 31	-2.95	0.20	-2.99	0.20
C 32	-1.06	0.17	-1.09	0.17
C 33	-1.74	0.17	-1.76	0.18
C 34	-0.74	0.16	-0.78	0.17
C 35	-0.25	0.16	-0.29	0.17
C 36	-1.82	0.17	-1.82	0.18
C 37	-0.82	0.16	-0.87	0.17
D 38	-1.45	0.17	-1.46	0.17
D 39	-1.12	0.17	-1.15	0.17
D 40	-1.85	0.17	-1.86	0.18
D 41	-1.76	0.17	-1.79	0.18
D 42	0.67	0.17	0.64	0.17
D 43	-2.10	0.18	-2.21	0.18
D 44	-0.46	0.16	-0.48	0.17
D 45	-1.06	0.17	-1.09	0.17
D 46	1.29	0.18	1.35	0.19
D 47	-1.23	0.17	-1.26	0.17
D 48	-0.56	0.16	-0.59	0.17
D 49	-1.01	0.16	-1.06	0.17
D 50	1.09	0.18	1.08	0.18

SET 1 (All measurable persons included)			SET 2 (7 misfitting persons excluded)	
ITEM NAME	ITEM DIFFICULTY	STANDARD ERROR	ITEM DIFFICULTY	STANDARD ERROR
D 51	-1.28	0.17	-1.35	0.17
D 52	-1.36	0.17	-1.43	0.17
E 53	0.28	0.17	0.35	0.17
E 54	1.29	0.18	1.31	0.19
E 55	-1.50	0.17	-1.52	0.17
E 56	-2.13	0.18	-2.11	0.18
E 57	-0.85	0.16	-0.84	0.17
E 58	-0.96	0.16	-1.01	0.17
E 59	-2.64	0.19	-2.71	0.20
E 60	-1.65	0.17	-1.67	0.17
E 61	0.67	0.17	0.67	0.17
E 62	1.68	0.20	1.64	0.20
E 63	-0.85	0.16	-0.87	0.17
E 64	2.29	0.23	2.26	0.23
F 65	-2.36	0.19	-2.42	0.19
F 66	-0.38	0.16	-0.40	0.17
F 67	-0.69	0.16	-0.65	0.17
F 68	-0.53	0.16	-0.59	0.17
F 69	-0.04	0.16	-0.02	0.17
F 70	-1.14	0.17	-1.20	0.17
F 71	-0.30	0.16	-0.29	0.17
F 72	-1.17	0.17	-1.17	0.17
F 73	-2.20	0.18	-2.24	0.19
F 74	0.76	0.17	0.79	0.18
F 75	0.31	0.17	0.32	0.17
F 76	-0.06	0.16	-0.02	0.17
G 77	1.46	0.19	1.53	0.19
G 78	1.49	0.19	1.57	0.20
G 79	1.46	0.19	1.49	0.19
G 80	-1.53	0.17	-1.55	0.17
G 81	-2.91	0.20	-2.90	0.20
G 82	1.53	0.19	1.53	0.19
G 83	-2.26	0.18	-2.28	0.19
G 84	-1.17	0.17	-1.23	0.17
G 85	-1.79	0.17	-1.86	0.18
G 86	0.59	0.17	0.64	0.17
G 87	0.23	0.17	0.26	0.17
H 88	-2.01	0.18	-2.01	0.18
H 89	1.29	0.18	1.31	0.19
H 90	-0.85	0.16	-0.92	0.17
H 91	1.83	0.20	1.80	0.20
H 92	3.91	0.40	3.91	0.40
H 93	-1.68	0.17	-1.67	0.17
H 94	1.83	0.20	1.84	0.21
H 95	2.89	0.27	3.02	0.29
H 96	0.85	0.17	0.92	0.18
H 97	1.60	0.19	1.68	0.20
I 98	5.95	1.01	5.94	1.01
I 99	-0.43	0.16	-0.45	0.17
I100	1.79	0.20	1.80	0.20

SET 1 (All measurable persons included)			SET 2 (7 misfitting persons excluded)	
ITEM NAME	ITEM DIFFICULTY	STANDARD ERROR	ITEM DIFFICULTY	STANDARD ERROR
I101	2.62	0.25	2.66	0.25
I102	-0.35	0.16	-0.32	0.17
I103	-0.69	0.16	-0.62	0.17
I104	0.76	0.17	0.79	0.18
I105	2.19	0.22	2.26	0.23
I106	1.75	0.20	1.84	0.21
I107	0.20	0.17	0.24	0.17
I108	0.04	0.16	0.10	0.17
J109	5.95	1.01	5.94	1.01
J110	1.92	0.21	1.93	0.21
J111	4.82	0.59	5.24	0.72
J112	1.96	0.21	1.97	0.21
J113	-0.74	0.16	-0.73	0.17
J114	-1.85	0.17	-1.89	0.18
J115	1.42	0.19	1.49	0.19
J116	0.02	0.16	0.04	0.17
J117	1.09	0.18	1.17	0.18
J118	1.64	0.20	1.72	0.20
J119	0.26	0.17	0.29	0.17
K120	0.20	0.17	0.21	0.17
K121	0.79	0.17	0.85	0.18
K122	3.63	0.36	3.76	0.38
K123	1.35	0.19	1.42	0.19
K124	1.88	0.20	1.89	0.21
K125	1.35	0.19	1.38	0.19
K126	0.79	0.17	0.82	0.18
K127	0.48	0.17	0.52	0.17
K128	2.00	0.21	1.97	0.21
K129	1.06	0.18	1.14	0.18
L130	3.77	0.38	3.76	0.38
L131	1.57	0.19	1.64	0.20
L132	-0.53	0.16	-0.51	0.17
L133	2.34	0.23	2.42	0.24
L134	2.09	0.22	2.11	0.22
L135	0.42	0.17	0.43	0.17
L136	2.62	0.25	2.73	0.26
L137	0.94	0.18	0.95	0.18
L138	3.21	0.30	3.20	0.30
L139	3.04	0.29	3.02	0.29
L140	-0.38	0.16	-0.37	0.17
L141	0.79	0.17	0.73	0.17
Items calibrated on 243 persons Mean item difficulty = 0.00 SD item difficulty = 1.91 Difficulty range: -4.08 to 5.95			Items calibrated on 236 persons Mean item difficulty = 0.00 SD item difficulty = 1.94 Difficulty range: -4.16 to 5.94	

F.3 Cloze-Type Test (Tanzanian Group): Observed ICCs & Departures from Expectation

	ITEM CHARACTERISTIC CURVE						DEPARTURE FROM EXPECTED ICC					
ITEM NAME	SUBGROUP 1	2	3	4	5	6	SUBGROUP 1	2	3	4	5	6
A 1	0.61	0.95	0.95	0.95	1.00	1.00	-0.03	0.06	-0.01	-0.03	0.01	0.00
A 2	0.15	0.54	0.97	1.00	1.00	1.00	-0.17	-0.10	0.14	0.08	0.03	0.01
A 3	0.15	0.51	0.82	0.90	0.97	0.95	-0.08	-0.01	0.07	0.02	0.03	-0.03
A 4	0.02	0.00	0.00	0.00	0.05	0.11	0.02	-0.00	-0.01	-0.02	0.01	-0.01
A 5	0.17	0.10	0.21	0.25	0.50	0.73	0.15	0.03	0.02	-0.10	-0.06	-0.04
A 6	0.22	0.41	0.41	0.62	0.87	1.00	0.11	0.10	-0.14	-0.12	0.00	0.05
A 7	0.02	0.08	0.28	0.47	0.37	0.27	0.01	0.03	0.16	0.23	-0.06	-0.40
A 8	0.49	0.74	0.77	0.67	0.75	0.95	0.25	0.19	-0.00	-0.21	-0.20	-0.03
A 9	0.39	0.85	0.87	0.92	0.97	0.95	-0.02	0.11	-0.02	-0.02	-0.00	-0.05
A 10	0.37	0.87	0.92	0.95	1.00	0.97	-0.10	0.09	0.01	-0.01	0.02	-0.02
A 11	0.46	0.85	0.90	1.00	1.00	1.00	-0.06	0.03	-0.03	0.03	0.01	0.01
B 12	0.12	0.15	0.67	0.90	0.95	0.97	-0.01	-0.22	0.05	0.11	0.05	0.01
B 13	0.71	0.87	0.72	0.77	0.82	0.68	0.41	0.25	-0.10	-0.14	-0.14	-0.31
B 14	0.17	0.21	0.38	0.77	0.82	1.00	0.08	-0.07	-0.12	0.07	-0.02	0.06
B 15	0.02	0.05	0.28	0.67	0.90	1.00	-0.04	-0.14	-0.11	0.07	0.12	0.09
B 16	0.05	0.13	0.13	0.17	0.45	0.68	0.03	0.08	-0.00	-0.09	-0.00	-0.02
B 17	0.10	0.10	0.18	0.47	0.67	0.78	0.07	-0.01	-0.07	0.03	0.02	-0.05
B 18	0.29	0.79	0.79	0.92	0.95	0.97	-0.04	0.13	-0.05	-0.00	-0.02	-0.02
B 19	0.20	0.51	0.72	0.80	0.85	0.97	0.02	0.06	0.03	-0.04	-0.08	0.00
B 20	0.00	0.03	0.03	0.12	0.12	0.35	-0.00	0.01	-0.01	0.05	-0.04	-0.01
B 21	0.12	0.13	0.46	0.90	0.90	0.97	0.02	-0.17	-0.08	0.17	0.03	0.03
B 22	0.39	0.64	0.69	0.80	0.87	1.00	0.14	0.08	-0.09	-0.09	-0.08	0.02
B 23	0.10	0.33	0.72	0.97	0.97	1.00	-0.09	-0.13	0.01	0.13	0.05	0.03
B 24	0.29	0.49	0.62	0.52	0.80	0.97	0.17	0.14	0.02	-0.25	-0.09	0.02
C 25	0.59	0.95	0.97	0.95	0.95	0.89	0.02	0.10	0.03	-0.02	-0.04	-0.10
C 26	0.02	0.31	0.56	0.95	0.95	0.97	-0.11	-0.06	-0.05	0.16	0.05	0.01
C 27	0.05	0.00	0.28	0.70	0.72	0.84	0.01	-0.14	-0.03	0.18	0.01	-0.03
C 28	0.10	0.13	0.05	0.17	0.37	0.73	0.08	0.08	-0.07	-0.07	-0.06	0.05
C 29	0.05	0.13	0.44	0.77	0.82	1.00	-0.03	-0.11	-0.03	0.10	-0.00	0.07
C 30	0.10	0.41	0.72	0.90	1.00	1.00	-0.09	-0.06	0.01	0.05	0.07	0.03
C 31	0.44	0.79	0.85	0.95	0.95	0.89	0.05	0.08	-0.03	0.01	-0.03	-0.10
C 32	0.12	0.28	0.59	0.75	0.85	0.81	0.03	-0.00	0.07	0.03	-0.01	-0.13
C 33	0.05	0.38	0.87	0.90	0.82	0.97	-0.12	-0.05	0.19	0.07	-0.10	0.00
C 34	0.05	0.23	0.54	0.67	0.72	0.92	-0.02	0.00	0.09	0.02	-0.09	-0.00
C 35	0.00	0.21	0.33	0.67	0.72	0.73	-0.05	0.05	0.00	0.14	-0.01	-0.15
C 36	0.10	0.49	0.69	0.87	0.92	0.97	-0.08	0.04	0.00	0.03	-0.00	0.00
C 37	0.00	0.18	0.44	0.75	0.90	0.95	-0.08	-0.06	-0.03	0.08	0.07	0.02
D 38	0.10	0.33	0.62	0.85	0.87	0.97	-0.03	-0.03	0.00	0.06	-0.02	0.01
D 39	0.22	0.38	0.54	0.67	0.72	0.92	0.12	0.09	0.00	-0.05	-0.14	-0.03
D 40	0.00	0.44	0.82	0.90	0.95	0.97	-0.18	-0.02	0.12	0.06	0.02	0.00
D 41	0.32	0.67	0.56	0.72	0.80	0.95	0.14	0.23	-0.12	-0.11	-0.12	-0.02
D 42	0.00	0.03	0.10	0.45	0.52	0.73	-0.02	-0.04	-0.06	0.13	0.00	-0.02
D 43	0.34	0.54	0.72	0.90	0.90	0.95	0.11	-0.00	-0.05	0.01	-0.05	-0.03
D 44	0.05	0.13	0.49	0.60	0.70	0.89	-0.01	-0.05	0.11	0.02	-0.07	-0.01
D 45	0.07	0.31	0.46	0.80	0.85	0.92	-0.02	0.02	-0.06	0.08	-0.01	-0.02
D 46	0.05	0.00	0.08	0.20	0.32	0.62	0.04	-0.03	-0.01	0.01	-0.03	0.02
D 47	0.10	0.33	0.64	0.70	0.85	0.95	-0.01	0.01	0.08	-0.05	-0.03	-0.00
D 48	0.34	0.41	0.38	0.52	0.60	0.68	0.28	0.21	-0.01	-0.08	-0.18	-0.23
D 49	0.05	0.31	0.67	0.62	0.82	0.92	-0.04	0.03	0.15	-0.09	-0.03	-0.02
D 50	0.02	0.08	0.15	0.15	0.32	0.76	0.01	0.03	0.04	-0.08	-0.09	0.10

	ITEM CHARACTERISTIC CURVE						DEPARTURE FROM EXPECTED ICC					
ITEM NAME	SUBGROUP 1	2	3	4	5	6	SUBGROUP 1	2	3	4	5	6
D 51	0.17	0.49	0.51	0.60	0.90	0.97	0.05	0.15	-0.07	-0.17	0.02	0.02
D 52	0.27	0.64	0.69	0.57	0.72	0.81	0.14	0.28	0.09	-0.21	-0.17	-0.15
E 53	0.02	0.05	0.18	0.40	0.55	0.89	-0.00	-0.04	-0.03	0.02	-0.04	0.09
E 54	0.00	0.05	0.08	0.15	0.35	0.68	-0.01	0.02	-0.01	-0.04	-0.01	0.07
E 55	0.10	0.23	0.59	0.87	1.00	1.00	-0.04	-0.15	-0.03	0.08	0.10	0.04
E 56	0.20	0.44	0.82	0.90	0.95	0.97	-0.02	-0.08	0.07	0.02	0.01	-0.01
E 57	0.05	0.13	0.38	0.70	0.95	0.97	-0.03	-0.11	-0.07	0.03	0.13	0.05
E 58	0.07	0.13	0.64	0.80	0.80	0.89	-0.02	-0.14	0.14	0.10	-0.05	-0.05
E 59	0.46	0.72	0.79	0.85	0.92	0.95	0.13	0.06	-0.05	-0.08	-0.04	-0.04
E 60	0.20	0.44	0.72	0.82	0.85	0.89	0.04	0.02	0.06	0.01	-0.06	-0.07
E 61	0.00	0.03	0.13	0.45	0.47	0.73	-0.02	-0.04	-0.03	0.14	-0.04	-0.01
E 62	0.00	0.00	0.00	0.12	0.32	0.62	-0.01	-0.03	-0.07	-0.02	0.03	0.09
E 63	0.05	0.15	0.49	0.72	0.82	0.97	-0.03	-0.09	0.02	0.05	-0.00	0.04
E 64	0.00	0.00	0.03	0.05	0.17	0.46	-0.00	-0.01	-0.01	-0.03	-0.01	0.08
F 65	0.39	0.59	0.85	0.72	0.97	0.97	0.12	-0.00	0.05	-0.18	0.02	-0.01
F 66	0.02	0.03	0.23	0.55	0.95	1.00	-0.03	-0.14	-0.12	-0.01	0.20	0.11
F 67	0.00	0.08	0.41	0.67	0.87	0.97	-0.06	-0.13	-0.00	0.05	0.08	0.06
F 68	0.10	0.13	0.44	0.70	0.77	0.81	0.04	-0.07	0.04	0.09	-0.01	-0.10
F 69	0.00	0.10	0.23	0.52	0.67	0.89	-0.04	-0.02	-0.04	0.06	0.00	0.04
F 70	0.10	0.41	0.56	0.65	0.85	0.95	-0.01	0.10	0.02	-0.09	-0.02	-0.00
F 71	0.02	0.10	0.46	0.57	0.67	0.84	-0.02	-0.05	0.13	0.04	-0.06	-0.04
F 72	0.05	0.21	0.62	0.80	0.85	0.97	-0.05	-0.10	0.07	0.07	-0.02	0.03
F 73	0.17	0.51	0.79	0.92	0.97	1.00	-0.07	-0.04	0.02	0.04	0.03	0.02
F 74	0.02	0.18	0.28	0.30	0.37	0.54	0.01	0.12	0.14	0.02	-0.11	-0.18
F 75	0.00	0.05	0.08	0.37	0.70	0.92	-0.03	-0.04	-0.14	-0.01	0.10	0.12
F 76	0.05	0.13	0.18	0.47	0.72	0.86	0.01	0.01	-0.09	0.01	0.05	0.02
G 77	0.07	0.08	0.03	0.22	0.17	0.57	0.07	0.05	-0.05	0.06	-0.14	0.01
G 78	0.02	0.00	0.00	0.10	0.25	0.76	0.02	-0.03	-0.07	-0.06	-0.06	0.21
G 79	0.00	0.05	0.03	0.20	0.37	0.51	-0.01	0.02	-0.05	0.03	0.05	-0.05
G 80	0.10	0.21	0.69	0.95	0.87	1.00	-0.04	-0.18	0.06	0.15	-0.03	0.04
G 81	0.27	0.69	0.90	1.00	0.97	1.00	-0.10	-0.01	0.03	0.06	0.00	0.01
G 82	0.00	0.03	0.05	0.05	0.30	0.73	-0.01	-0.00	-0.02	-0.11	-0.01	0.17
G 83	0.22	0.46	0.85	0.90	0.97	1.00	-0.03	-0.10	0.07	0.01	0.02	0.02
G 84	0.15	0.41	0.62	0.77	0.67	0.92	0.04	0.10	0.06	0.03	-0.20	-0.03
G 85	0.12	0.44	0.72	0.87	0.92	1.00	-0.06	-0.02	0.02	0.03	-0.00	0.03
G 86	0.02	0.00	0.13	0.32	0.52	0.84	0.01	-0.07	-0.04	0.01	0.00	0.09
G 87	0.00	0.00	0.08	0.40	0.72	0.97	-0.03	-0.09	-0.15	-0.00	0.11	0.16
H 88	0.10	0.54	0.77	0.87	0.92	1.00	-0.11	0.04	0.04	0.01	-0.01	0.02
H 89	0.00	0.00	0.15	0.15	0.32	0.68	-0.01	-0.04	0.06	-0.04	-0.04	0.07
H 90	0.00	0.18	0.36	0.82	0.97	0.92	-0.08	-0.07	-0.12	0.14	0.14	-0.01
H 91	0.02	0.00	0.13	0.10	0.20	0.51	0.02	-0.02	0.07	-0.03	-0.06	0.02
H 92	0.00	0.03	0.03	0.00	0.02	0.11	-0.00	0.02	0.02	-0.02	-0.02	-0.01
H 93	0.00	0.41	0.69	0.87	0.97	0.97	-0.16	-0.00	0.03	0.06	0.06	0.01
H 94	0.02	0.05	0.13	0.15	0.27	0.30	0.02	0.03	0.07	0.03	0.02	-0.18
H 95	0.00	0.00	0.00	0.05	0.15	0.19	-0.00	-0.01	-0.02	0.01	0.06	-0.04
H 96	0.00	0.03	0.08	0.20	0.45	0.86	-0.01	-0.03	-0.05	-0.06	-0.00	0.17
H 97	0.00	0.03	0.03	0.12	0.32	0.54	-0.01	0.00	-0.04	-0.02	0.04	0.02
I 98	0.00	0.00	0.00	0.00	0.02	0.00	-0.00	-0.00	-0.00	-0.00	0.02	-0.02
I 99	0.00	0.31	0.33	0.55	0.75	0.89	-0.05	0.13	-0.03	-0.03	-0.01	-0.00
I100	0.00	0.03	0.00	0.05	0.20	0.70	-0.01	0.00	-0.06	-0.08	-0.06	0.21

ITEM NAME	ITEM CHARACTERISTIC CURVE SUBGROUP 1	2	3	4	5	6	DEPARTURE FROM EXPECTED ICC SUBGROUP 1	2	3	4	5	6
I101	0.00	0.05	0.03	0.05	0.15	0.24	-0.00	0.04	0.00	-0.01	0.02	-0.05
I102	0.00	0.08	0.28	0.62	0.80	0.92	-0.05	-0.08	-0.05	0.08	0.06	0.04
I103	0.05	0.13	0.36	0.70	0.75	1.00	-0.01	-0.07	-0.05	0.08	-0.04	0.09
I104	0.00	0.03	0.13	0.27	0.42	0.86	-0.02	-0.03	-0.02	-0.01	-0.06	0.15
I105	0.00	0.00	0.03	0.10	0.07	0.51	-0.00	-0.01	-0.01	0.02	-0.11	0.13
I106	0.00	0.00	0.00	0.10	0.20	0.65	-0.01	-0.02	-0.06	-0.02	-0.05	0.17
I107	0.12	0.13	0.13	0.30	0.55	0.97	0.09	0.03	-0.10	-0.11	-0.07	0.16
I108	0.02	0.08	0.26	0.45	0.67	0.84	-0.01	-0.03	0.00	0.01	0.02	0.00
J109	0.00	0.00	0.00	0.00	0.00	0.03	-0.00	-0.00	-0.00	-0.00	-0.01	0.01
J110	0.00	0.00	0.03	0.10	0.25	0.51	-0.01	-0.02	-0.03	-0.01	0.01	0.05
J111	0.00	0.03	0.00	0.00	0.00	0.03	-0.00	0.02	-0.00	-0.00	-0.01	-0.01
J112	0.00	0.00	0.05	0.17	0.22	0.41	-0.01	-0.02	0.00	0.07	-0.00	-0.04
J113	0.02	0.21	0.49	0.62	0.80	0.95	-0.04	-0.01	0.06	-0.02	-0.01	0.03
J114	0.05	0.36	0.82	0.87	1.00	1.00	-0.14	-0.10	0.12	0.03	0.07	0.03
J115	0.00	0.03	0.18	0.20	0.37	0.38	-0.01	-0.00	0.10	0.03	0.05	-0.19
J116	0.00	0.00	0.21	0.35	0.90	0.92	-0.03	-0.12	-0.06	-0.11	0.24	0.08
J117	0.00	0.13	0.13	0.30	0.30	0.54	-0.01	0.09	0.02	0.09	-0.09	-0.10
J118	0.00	0.03	0.03	0.15	0.45	0.35	-0.01	0.00	-0.04	0.01	0.18	-0.16
J119	0.00	0.23	0.26	0.45	0.50	0.70	-0.03	0.14	0.04	0.06	-0.11	-0.10
K120	0.02	0.21	0.36	0.30	0.55	0.78	-0.00	0.11	0.13	-0.11	-0.07	-0.04
K121	0.00	0.05	0.21	0.15	0.62	0.62	-0.02	-0.00	0.07	-0.12	0.16	-0.08
K122	0.00	0.00	0.03	0.02	0.12	0.03	-0.00	-0.00	0.02	0.00	0.08	-0.10
K123	0.02	0.00	0.21	0.12	0.42	0.43	0.02	-0.03	0.12	-0.05	0.09	-0.15
K124	0.00	0.03	0.03	0.02	0.27	0.57	-0.01	0.01	-0.03	-0.09	0.03	0.10
K125	0.00	0.05	0.08	0.20	0.45	0.46	-0.01	0.02	-0.01	0.02	0.10	-0.13
K126	0.00	0.00	0.08	0.22	0.47	0.92	-0.02	-0.06	-0.06	-0.05	-0.00	0.21
K127	0.00	0.05	0.21	0.37	0.57	0.73	-0.02	-0.02	0.02	0.03	0.02	-0.04
K128	0.00	0.00	0.00	0.10	0.37	0.38	-0.01	-0.02	-0.05	-0.01	0.15	-0.07
K129	0.00	0.00	0.15	0.25	0.45	0.57	-0.01	-0.04	0.05	0.03	0.05	-0.08
L130	0.00	0.05	0.00	0.02	0.07	0.05	-0.00	0.05	-0.01	0.00	0.03	-0.08
L131	0.00	0.00	0.03	0.10	0.32	0.62	-0.01	-0.03	-0.04	-0.05	0.03	0.09
L132	0.00	0.08	0.38	0.52	0.90	1.00	-0.06	-0.11	0.00	-0.06	0.13	0.10
L133	0.00	0.00	0.05	0.00	0.25	0.32	-0.00	-0.01	0.02	-0.07	0.09	-0.02
L134	0.00	0.00	0.03	0.05	0.15	0.57	-0.00	-0.02	-0.02	-0.05	-0.05	0.15
L135	0.00	0.05	0.18	0.22	0.77	0.78	-0.02	-0.03	-0.01	-0.14	0.20	0.00
L136	0.00	0.03	0.03	0.10	0.15	0.19	-0.00	0.02	0.00	0.05	0.03	-0.10
L137	0.00	0.05	0.08	0.07	0.45	0.95	-0.01	0.00	-0.05	-0.18	0.00	0.26
L138	0.00	0.03	0.00	0.00	0.02	0.30	-0.00	0.02	-0.02	-0.04	-0.06	0.09
L139	0.00	0.00	0.00	0.02	0.05	0.32	-0.00	-0.01	-0.02	-0.02	-0.04	0.09
L140	0.00	0.10	0.26	0.55	0.92	0.92	-0.05	-0.06	-0.09	-0.01	0.18	0.03
L141	0.00	0.18	0.13	0.30	0.57	0.57	-0.02	0.12	-0.02	0.00	0.08	-0.16

GROUP	SCORE RANGE	MEAN ABILITY	NO. IN SUBGROUP
1	1 - 22	-3.52	41
2	23 - 42	-2.04	39
3	43 - 59	-1.01	39
4	60 - 77	-0.14	40
5	78 - 97	0.73	40
6	98 - 140	1.78	37

N = 236

F.4 Cloze-Type Test (Tanzanian Group): Item Fit Statistics

(Items ordered by total fit-t; 7 misfitting persons omitted)

ITEM NAME	ITEM DIFFIC.	BETWEEN GROUP FIT-T	TOTAL FIT-T	RASCH DISCRIM. INDEX
G 87	0.26	2.84	-4.86	1.43
F 66	-0.40	3.26	-4.81	1.50
B 15	-0.56	2.48	-4.55	1.45
J116	0.04	3.13	-4.41	1.41
J114	-1.89	2.15	-3.92	1.36
A 2	-2.63	2.82	-3.83	1.33
H 90	-0.92	2.58	-3.82	1.38
F 67	-0.65	1.57	-3.73	1.39
K126	0.82	1.96	-3.72	1.29
F 75	0.32	1.67	-3.64	1.31
L132	-0.51	2.17	-3.59	1.41
L137	0.95	3.02	-3.48	1.28
L140	-0.37	1.82	-3.40	1.34
D 40	-1.86	2.10	-3.36	1.31
E 55	-1.52	1.80	-3.32	1.34
C 26	-1.49	2.01	-3.29	1.35
H 93	-1.67	1.63	-3.26	1.32
B 23	-1.89	1.87	-3.22	1.33
E 57	-0.84	1.41	-3.12	1.34
G 80	-1.55	2.19	-2.87	1.27
I102	-0.32	0.76	-2.77	1.28
C 30	-1.92	0.95	-2.76	1.28
C 37	-0.87	0.87	-2.75	1.31
B 12	-1.49	1.94	-2.74	1.27
H 96	0.92	1.05	-2.61	1.22
C 29	-0.87	1.08	-2.58	1.28
G 78	1.57	2.17	-2.56	1.19
B 21	-1.17	2.06	-2.42	1.24
I100	1.80	1.91	-2.22	1.20
G 86	0.64	0.19	-2.02	1.16
G 82	1.53	1.25	-1.86	1.17
G 85	-1.86	-0.76	-1.79	1.14
E 63	-0.87	-0.16	-1.77	1.20
F 72	-1.17	0.36	-1.75	1.20
I103	-0.62	0.79	-1.68	1.21
L135	0.43	1.77	-1.67	1.17
G 81	-2.90	0.30	-1.63	1.16
C 33	-1.76	2.73	-1.62	1.14
E 53	0.35	-0.41	-1.62	1.12
I104	0.79	0.55	-1.61	1.17
E 62	1.64	0.47	-1.61	1.16
I106	1.84	1.14	-1.56	1.19
G 83	-2.28	-0.14	-1.56	1.13
C 27	-0.21	1.94	-1.55	1.13
H 88	-2.01	0.12	-1.49	1.15
F 69	-0.02	-0.50	-1.48	1.16
L131	1.64	0.11	-1.47	1.15

(Items ordered by total fit-t; 7 misfitting persons omitted)

ITEM NAME	ITEM DIFFIC.	BETWEEN GROUP FIT-T	TOTAL FIT-T	RASCH DISCRIM. INDEX
C 36	-1.82	-0.79	-1.43	1.11
A 10	-3.34	0.78	-1.34	1.06
K124	1.89	0.49	-1.32	1.12
I105	2.26	0.75	-1.24	1.11
F 73	-2.24	-0.38	-1.22	1.17
L134	2.11	0.64	-1.22	1.16
A 3	-2.14	0.32	-1.15	1.10
E 56	-2.11	-0.70	-1.09	1.09
D 42	0.64	0.62	-0.97	1.09
I108	0.10	-2.13	-0.97	1.06
E 64	2.26	-0.79	-0.87	1.10
F 76	-0.02	-0.75	-0.82	1.05
D 38	-1.46	-1.07	-0.81	1.10
H 97	1.68	-1.15	-0.80	1.07
E 54	1.31	-1.02	-0.71	1.06
E 58	-1.01	1.66	-0.69	1.04
B 14	-1.03	1.45	-0.67	1.02
L139	3.02	-0.07	-0.65	1.13
J113	-0.73	-0.89	-0.62	1.11
J110	1.93	-0.98	-0.51	1.10
D 45	-1.09	-0.70	-0.48	1.02
I107	0.24	3.58	-0.46	0.96
H 89	1.31	0.25	-0.40	1.07
L138	3.20	1.18	-0.33	1.09
K127	0.52	-1.04	-0.26	1.03
D 50	1.08	0.67	-0.24	0.98
E 61	0.67	0.53	-0.21	1.06
A 4	3.91	4.29	-0.21	0.89
G 79	1.49	-0.24	-0.11	0.99
A 11	-3.58	-0.38	-0.09	1.09
K128	1.97	1.20	-0.01	1.05
L133	2.42	0.72	0.01	1.03
B 20	2.37	-0.73	0.03	0.97
H 92	3.91	1.48	0.09	0.88
H 95	3.02	-0.56	0.14	1.01
K121	0.85	1.55	0.21	0.99
A 1	-4.16	0.03	0.23	1.01
J111	5.24	4.44	0.26	0.69
B 18	-2.75	0.35	0.26	0.96
K129	1.14	0.13	0.29	0.99
A 9	-3.11	1.96	0.29	0.91
J109	5.94	-2.22	0.30	1.12
D 46	1.35	1.13	0.31	0.98
J112	1.97	-0.52	0.31	0.98
I 98	5.94	-0.33	0.33	0.89
C 28	0.98	4.19	0.38	0.83
D 47	-1.26	-1.07	0.40	0.95
K122	3.76	1.50	0.45	0.84
I101	2.66	1.03	0.51	0.90
L130	3.76	4.20	0.52	0.75
C 34	-0.78	-0.12	0.54	0.95
I 99	-0.45	0.90	0.55	0.97

(Items ordered by total fit-t; 7 misfitting persons omitted)

ITEM NAME	ITEM DIFFIC.	BETWEEN GROUP FIT-T	TOTAL FIT-T	RASCH DISCRIM. INDEX
F 68	-0.59	1.03	0.68	0.87
H 91	1.80	0.97	0.69	0.95
D 44	-0.48	-0.14	0.73	0.97
B 16	0.92	1.37	0.74	0.85
B 17	0.10	0.98	0.75	0.88
F 71	-0.29	0.46	0.86	0.96
D 49	-1.06	0.76	0.90	0.94
F 70	-1.20	-0.17	0.92	0.88
C 25	-3.80	8.69	0.94	0.75
J118	1.72	1.72	0.94	0.96
L136	2.73	0.19	1.03	0.88
B 19	-1.82	0.11	1.13	0.85
K125	1.38	0.40	1.29	0.93
G 77	1.53	4.26	1.40	0.81
C 32	-1.09	2.01	1.47	0.77
C 35	-0.29	2.17	1.60	0.87
C 31	-2.99	5.18	1.61	0.77
D 51	-1.35	1.90	1.74	0.80
D 43	-2.21	0.88	1.76	0.79
E 60	-1.67	1.40	1.81	0.77
J115	1.49	1.88	1.86	0.83
J117	1.17	2.03	1.89	0.80
K123	1.42	2.42	1.93	0.86
A 6	-1.23	2.40	2.12	0.79
F 65	-2.42	2.90	2.19	0.78
H 94	1.84	2.10	2.20	0.75
L141	0.73	2.50	2.28	0.82
A 5	0.49	5.41	2.35	0.63
K120	0.21	1.81	2.49	0.77
G 84	-1.23	2.75	2.67	0.65
J119	0.29	2.38	2.70	0.74
E 59	-2.71	2.50	3.04	0.69
B 22	-2.28	2.57	3.49	0.70
D 39	-1.15	2.42	3.53	0.55
F 74	0.79	3.54	3.61	0.60
D 41	-1.79	4.11	4.28	0.48
B 24	-1.40	4.32	4.38	0.48
A 7	1.01	5.80	5.29	0.50
A 8	-2.24	7.16	5.60	0.28
D 52	-1.43	6.62	6.34	0.14
D 48	-0.59	8.45	7.73	-0.18
B 13	-2.56	14.81	7.80	-0.17

TOTAL FIT-T:
Mean = -0.28
SD = 2.38
Range = -4.86 to 7.80

BETWEEN-GROUP FIT-T:
Mean = 1.53
SD = 2.18
Range = -2.22 to 14.81

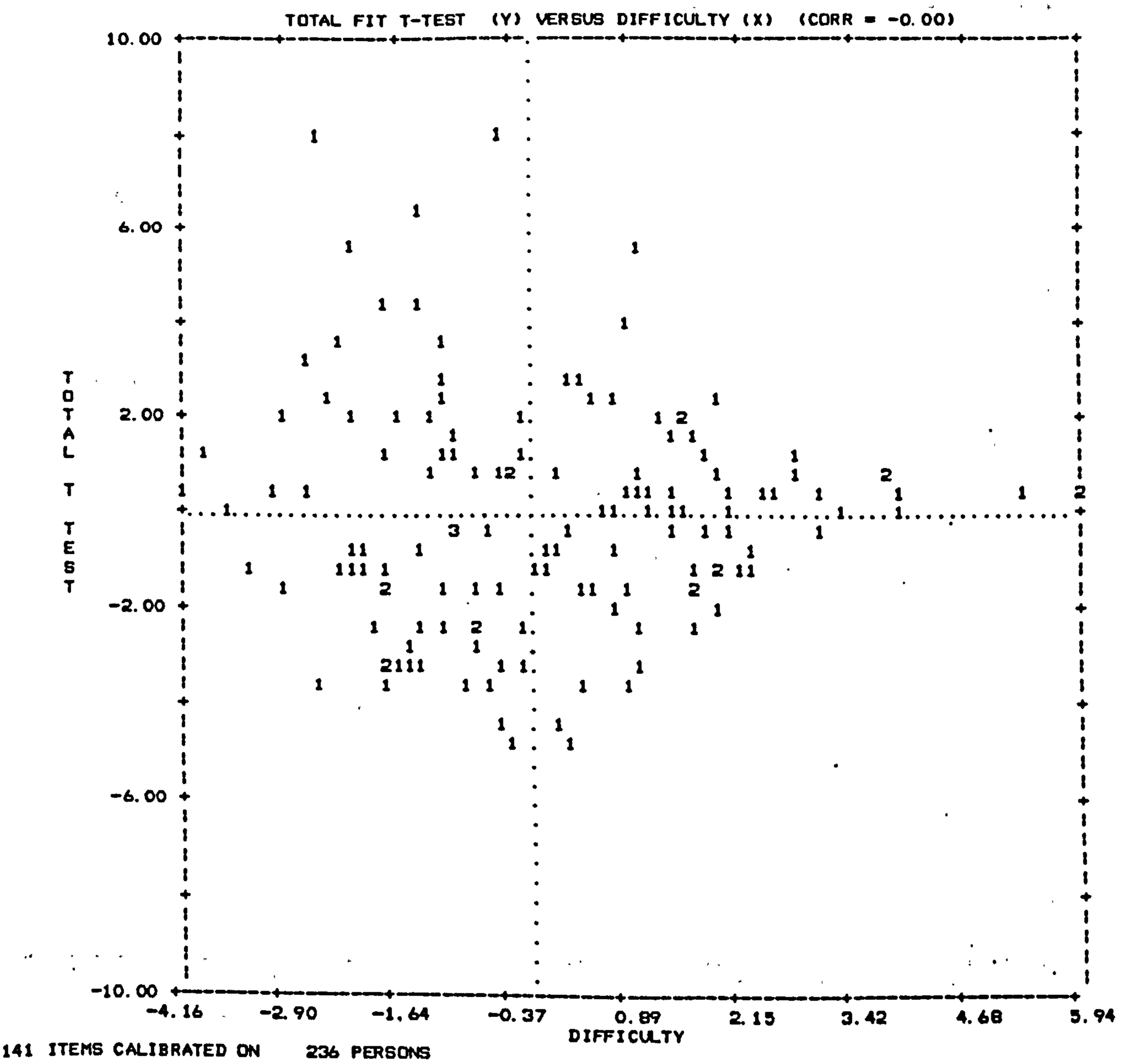
DISCRIM. INDEX:
Range = -1.17 to 1.50

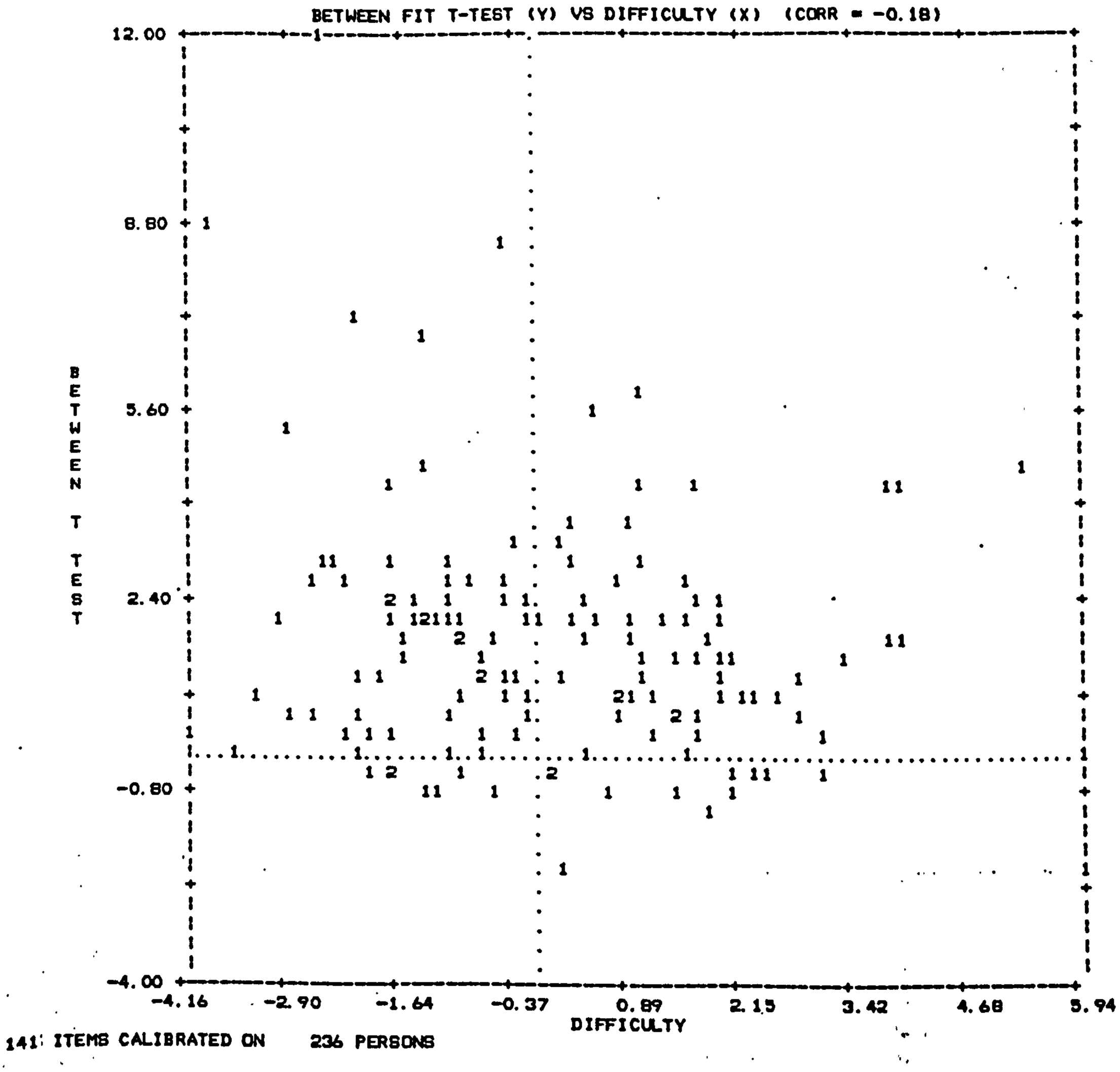
F.5 Person Statistics and Standardized Residuals for Misfitting Persons (Tanzanian Group)

Person No.	Raw Score	Abil. Estim.	Total Fit-t	Standardized Residuals (letters refer to passages)
97	74	0.07	2.92	**A :** 0 0 0 0 0 −1 0 0 0 0 0 **B :** 0 0 0 0 0 0 0 0 0 0 0 0 0 −1 **C :** 0 −2 −1 0 0 0 0 0 −2 −1 0 0 −1 **D :** 0 0 0 0 0 0 0 −1 0 −1 −1 0 0 −1 −2 **E :** 0 1 0 0 0 −1 0 0 0 0 −1 0 **F :** −3 −1 0 0 −1 −1 0 0 −3 0 1 0 **G :** 0 0 0 0 0 0 0 0 0 0 0 0 **H :** 0 0 −1 0 0 0 2 0 0 2 **I :** 0 0 0 3 0 0 1 0 2 1 0 **J :** 0 0 0 0 −1 0 1 0 1 2 0 **K :** 1 1 0 1 2 0 1 1 0 1 **L :** 0 0 0 0 2 0 0 1 0 0 −1 0
52	91	0.88	2.76	**A :** 0 0 0 0 −1 −2 0 0 −6 0 0 **B :** 0 0 0 0 0 −1 0 −3 2 0 0 0 −2 **C :** 0 0 0 0 0 0 0 0 0 0 −1 0 0 **D :** 0 0 0 0 −1 0 −1 0 1 −2 −2 −2 0 0 0 **E :** 0 0 0 0 0 0 0 0 0 0 0 0 **F :** 0 0 0 −2 0 0 0 0 0 0 0 0 **G :** 1 1 1 0 0 1 0 −2 0 0 0 **H :** 0 0 −2 0 0 0 0 0 0 1 **I :** 0 −1 0 0 0 0 −1 0 1 0 −1 **J :** 0 1 7 1 0 0 1 0 1 1 0 **K :** −1 −1 0 1 0 1 0 −1 0 0 **L :** 0 1 0 2 0 0 0 0 0 0 0 −1
166	40	−1.52	2.75	**A :** 0 −1 −1 0 0 0 0 −1 0 0 −2 **B :** 0 −1 1 0 0 0 −1 0 0 0 0 −1 0 **C :** 0 0 0 0 0 0 0 1 0 0 0 0 0 **D :** 0 1 0 −1 0 −1 0 0 4 1 1 0 0 0 0 **E :** 0 0 1 0 0 0 −1 −1 0 0 1 0 **F :** 0 0 1 0 2 0 0 1 −1 3 0 2 **G :** 0 0 0 −1 0 0 0 0 −1 0 0 **H :** 0 4 1 0 0 0 0 0 3 0 **I :** 0 0 0 0 0 1 3 6 0 2 2 **J :** 0 0 0 0 1 −1 0 0 0 0 2 **K :** 0 0 0 0 0 0 0 0 0 0 **L :** 0 0 0 0 0 0 0 0 0 0 0 0
171	38	−1.62	2.57	**A :** 0 0 0 0 0 0 0 0 −1 0 0 **B :** 1 −1 0 1 0 0 0 0 0 0 0 0 0 **C :** −2 0 0 0 0 0 −1 0 −1 0 0 −1 0 **D :** 1 1 −1 0 0 −1 0 1 0 0 1 0 0 1 0 **E :** 2 0 0 0 0 0 −1 −1 0 0 0 0 **F :** −1 0 1 0 0 0 0 0 0 3 0 2 **G :** 0 4 0 0 0 0 −1 0 −1 3 0 **H :** 0 0 0 0 0 −1 0 0 3 0 **I :** 0 0 0 0 1 1 0 0 0 0 0 **J :** 0 0 0 0 1 0 0 0 3 0 2 **K :** 2 0 0 0 0 0 0 0 0 3 **L :** 0 0 0 0 0 0 0 0 0 0 0 0 0

Person No.	Raw Score	Abil. Estim.	Total Fit-t	Standardized Residuals (letters refer to passages)
64	82	0.45	2.19	**A :** 0 -4 0 0 0 0 0 -3 -5 0 0 **B :** 0 0 -2 0 0 0 -4 -3 0 -2 0 0 0 **C :** 0 0 0 1 0 0 0 -2 0 0 -1 0 0 **D :** 0 -2 0 0 0 -3 0 0 0 0 0 -2 0 0 0 **E :** 0 1 0 0 0 0 0 0 0 0 0 0 0 **F :** 0 0 0 -1 0 0 0 0 0 0 -1 0 **G :** 1 1 0 0 0 0 0 0 0 0 1 0 **H :** 0 1 0 0 0 0 0 3 0 0 **I :** 0 0 1 0 0 0 1 0 1 0 0 **J :** 0 0 0 0 -1 -3 0 -1 0 0 -1 **K :** -1 1 0 0 0 0 1 1 0 1 **L :** 0 1 0 2 0 0 0 1 0 0 0 0
78	78	0.26	2.09	**A :** -8 0 0 0 0 -1 1 0 0 0 0 **B :** 0 -3 -1 0 0 -1 0 0 0 0 0 0 -2 **C :** 0 0 0 0 0 0 0 -1 0 0 -1 0 0 **D :** -2 -1 0 -2 1 0 0 0 1 0 -1 0 1 -2 0 **E :** 1 0 0 0 0 0 0 0 0 0 0 0 0 **F :** 0 0 0 0 0 -2 0 0 0 0 0 -1 **G :** 1 0 0 0 0 0 0 -2 0 1 0 **H :** 0 0 -1 0 0 0 0 3 1 1 **I :** 0 -1 0 0 -1 0 0 0 0 -1 0 **J :** 0 0 0 0 0 0 1 0 0 1 -1 **K :** 0 1 0 1 0 1 0 1 0 0 **L :** 0 0 0 0 0 1 3 0 0 0 0 0
118	62	-0.48	2.05	**A :** 0 0 0 0 0 -1 0 0 0 0 0 **B :** 0 0 -1 -1 0 0 0 0 0 0 -2 0 -1 **C :** 0 0 1 0 0 -2 0 0 0 -1 1 0 -1 **D :** 0 -1 0 0 0 -2 0 -1 0 0 -1 0 0 -1 -1 **E :** 1 0 -1 0 0 -1 0 0 1 0 -1 0 **F :** 0 0 -1 -1 0 0 0 -1 0 0 1 0 **G :** 0 0 2 0 0 0 0 0 -1 0 1 **H :** -2 0 -1 0 0 0 0 0 0 0 **I :** 0 1 0 0 1 0 0 3 0 0 1 **J :** 0 0 0 0 0 0 0 0 1 2 0 1 **K :** 0 1 7 0 0 0 0 1 0 2 **L :** 0 2 0 0 0 0 4 0 0 0 1 0

F.6 Cloze-Type Test (Tanzanian Group): Item Fit Statistics vs Item Difficulty





141 ITEMS CALIBRATED ON 236 PERSONS

APPENDIX G
CLOZE-TYPE DATA SUBSETS: RASCH DIFFICULTIES

G.1 Item Difficulty Estimates for High- & Low-Scoring Subgroups (Malaysian Data)

	200 Highest-Scoring Persons		200 Lowest-Scoring Persons	
ITEM NAME	ITEM DIFFIC.	STANDARD ERROR	ITEM DIFFIC.	STANDARD ERROR
A 1	-2.86	1.00	-3.61	0.24
A 3	-2.86	1.00	-2.52	0.19
A 4	1.32	0.17	2.27	0.35
A 5	0.45	0.22	1.64	0.27
A 6	-0.51	0.33	-1.54	0.17
A 7	0.06	0.26	-0.52	0.16
A 8	-2.16	0.71	-2.74	0.20
A 9	-2.16	0.71	-2.98	0.21
B 13	0.80	0.20	-2.74	0.20
B 14	-0.89	0.39	-0.22	0.17
B 16	1.21	0.18	0.71	0.20
B 17	1.35	0.17	0.19	0.18
B 18	-2.16	0.71	-2.59	0.19
B 20	-1.05	0.42	1.26	0.24
B 22	-2.86	1.00	-2.09	0.18
B 24	-1.05	0.42	-1.13	0.16
C 25	-2.16	0.71	-4.00	0.27
C 27	-0.08	0.27	1.79	0.29
C 28	2.36	0.15	2.89	0.46
C 29	-1.46	0.51	0.29	0.18
C 30	-1.75	0.58	-1.51	0.17
C 31	-1.75	0.58	-3.06	0.21
C 32	-1.23	0.45	-2.15	0.18
C 34	0.91	0.19	-2.31	0.18
C 35	0.35	0.23	-0.28	0.17
C 36	1.21	0.18	-0.63	0.16
C 37	-1.05	0.42	0.07	0.17
D 38	-2.16	0.71	-1.91	0.17
D 39	-0.51	0.33	-0.74	0.16
D 40	-0.00	0.26	-1.91	0.17
D 41	-2.86	1.00	-2.41	0.19
D 42	0.30	0.23	-0.14	0.17
D 43	-0.89	0.39	-0.79	0.16
D 44	0.45	0.22	-1.82	0.17
D 45	-1.46	0.51	-1.21	0.16
D 46	-1.23	0.45	0.79	0.20
D 47	-1.23	0.45	-2.03	0.18
D 48	-0.23	0.29	-1.00	0.16
D 49	-0.75	0.36	0.01	0.17
D 50	-1.05	0.42	1.79	0.29

	200 Highest-Scoring Persons		200 Lowest-Scoring Persons	
ITEM NAME	ITEM DIFFIC.	STANDARD ERROR	ITEM DIFFIC.	STANDARD ERROR
D 51	-2.86	1.00	-1.51	0.17
D 52	-2.16	0.71	-2.00	0.18
E 53	-0.41	0.31	-0.20	0.17
E 54	1.87	0.15	0.79	0.20
E 55	-2.16	0.71	-1.37	0.16
E 56	-2.16	0.71	-1.08	0.16
E 57	-1.75	0.58	-1.54	0.17
E 58	-1.23	0.45	-1.32	0.16
E 59	1.89	0.15	2.70	0.42
E 60	-1.46	0.51	-1.76	0.17
E 61	0.06	0.26	0.35	0.18
E 62	0.55	0.21	1.79	0.29
E 63	-0.75	0.36	0.16	0.18
E 64	1.54	0.16	2.27	0.35
F 65	0.13	0.25	-1.56	0.17
F 66	0.55	0.21	1.32	0.24
F 67	1.24	0.17	-1.29	0.16
F 68	0.35	0.23	0.53	0.19
F 69	0.55	0.21	1.16	0.23
F 70	-0.62	0.34	-0.50	0.16
F 71	-1.23	0.45	-1.21	0.16
F 72	-0.89	0.39	-0.76	0.16
F 73	0.13	0.25	-2.34	0.18
F 74	-0.75	0.36	0.45	0.19
F 75	-2.16	0.71	0.64	0.20
F 76	0.45	0.22	1.71	0.28
G 77	0.24	0.24	1.32	0.24
G 78	0.80	0.20	2.54	0.39
G 79	0.59	0.21	1.79	0.29
G 80	-2.16	0.71	-1.13	0.16
G 81	-1.46	0.51	-1.94	0.17
G 82	2.53	0.15	1.96	0.31
G 83	-1.75	0.58	-1.16	0.16
G 84	-2.16	0.71	-1.97	0.17
G 87	-0.23	0.29	0.25	0.18
H 88	-1.46	0.51	-1.29	0.16
H 89	3.27	0.16	2.70	0.42
H 90	-0.15	0.28	-0.36	0.17
H 91	0.45	0.22	2.54	0.39
H 92	2.97	0.15	2.54	0.39
H 93	-2.86	1.00	-1.16	0.16
H 94	1.94	0.15	1.06	0.22
H 95	2.86	0.15	2.70	0.42
H 96	-0.89	0.39	0.42	0.19
H 97	0.59	0.21	0.32	0.18
I 98	4.19	0.19	3.82	0.72
I 99	0.06	0.26	-0.50	0.16
I100	0.19	0.24	3.41	0.59

	200 Highest-Scoring Persons		200 Lowest-Scoring Persons	
ITEM NAME	ITEM DIFFIC.	STANDARD ERROR	ITEM DIFFIC.	STANDARD ERROR
I101	2.88	0.15	2.40	0.37
I102	-1.75	0.58	-0.14	0.17
I103	-0.51	0.33	0.67	0.20
I104	1.94	0.15	3.82	0.72
I105	-0.08	0.27	2.27	0.35
I106	0.55	0.21	1.32	0.24
I107	-2.86	1.00	-0.14	0.17
I108	-1.05	0.42	-0.02	0.17
J110	1.72	0.16	3.12	0.51
J111	4.64	0.22	3.12	0.51
J112	2.42	0.15	2.27	0.35
J113	-0.23	0.29	-0.74	0.16
J114	-1.75	0.58	-1.65	0.17
J115	2.38	0.15	2.16	0.33
J116	-2.86	1.00	1.38	0.24
J117	1.05	0.18	0.39	0.19
J118	1.62	0.16	2.54	0.39
J119	0.19	0.24	0.49	0.19
K120	-0.51	0.33	-0.74	0.16
K121	0.72	0.20	0.64	0.20
K123	2.05	0.15	1.11	0.22
K125	0.94	0.19	0.75	0.20
K126	1.29	0.17	2.05	0.32
K127	-1.46	0.51	0.25	0.18
K128	1.72	0.16	3.82	0.72
K129	0.50	0.22	1.06	0.22
L130	-1.23	0.45	-0.82	0.16
L132	-2.86	1.00	0.64	0.20
L133	3.37	0.16	4.52	1.01
L135	-0.75	0.36	0.75	0.20
L136	3.55	0.16	2.70	0.42
L137	-1.05	0.42	1.50	0.26
L138	2.88	0.15	4.52	1.01
L140	-0.08	0.27	-0.63	0.16
L141	1.29	0.17	0.16	0.18

G.2 Item Difficulty Estimates from Separate Calibrations of Content Word & Structure Word Items (Malaysian Data)

Content Word Items

ITEM NAME	ITEM DIFFIC.	STANDARD ERROR
A 1	-4.29	0.21
A 9	-3.86	0.19
B 14	-0.94	0.11
B 16	0.06	0.10
B 17	0.07	0.10
B 18	-3.16	0.16
C 26	-2.62	0.14
C 27	0.03	0.10
C 28	1.87	0.11
C 29	-1.25	0.11
C 32	-2.66	0.14
C 35	-0.92	0.11
C 36	-0.95	0.11
D 44	-1.98	0.13
D 49	-0.86	0.11
D 50	-0.09	0.10
E 55	-1.93	0.12
E 61	-0.51	0.11
E 63	-0.88	0.11
E 64	0.72	0.10
F 65	-1.53	0.12
F 66	0.09	0.10
F 67	-1.42	0.12
F 68	-0.18	0.11
F 69	-0.13	0.10
F 74	-0.60	0.11
F 75	-0.89	0.11
G 78	1.02	0.11
G 79	0.52	0.10
G 81	-2.33	0.13
G 82	1.43	0.11
G 86	0.93	0.11
H 89	2.44	0.12
H 90	-0.93	0.11
H 91	0.49	0.10
H 92	2.18	0.12
H 94	0.75	0.10
H 95	2.04	0.12
H 96	-0.99	0.11
H 97	-0.56	0.11
I 98	3.50	0.16

Structure Word Items

ITEM NAME	ITEM DIFFIC.	STANDARD ERROR
A 2	-2.63	0.20
A 3	-1.89	0.17
A 4	2.45	0.10
A 5	1.86	0.10
A 6	-0.34	0.12
A 7	1.21	0.10
A 8	-1.98	0.17
A 10	-2.85	0.22
A 11	-2.01	0.17
B 12	-1.09	0.14
B 13	-0.37	0.12
B 15	-0.21	0.12
B 19	-1.05	0.14
B 20	1.44	0.10
B 21	-1.35	0.15
B 22	-1.54	0.15
B 23	-1.44	0.15
B 24	-0.37	0.12
C 25	-3.21	0.24
C 30	-1.11	0.14
C 31	-2.13	0.18
C 33	-2.90	0.22
C 34	-0.71	0.13
C 37	0.37	0.11
D 38	-1.42	0.15
D 39	-0.07	0.12
D 40	-0.66	0.13
D 41	-1.76	0.16
D 42	0.54	0.11
D 43	-0.25	0.12
D 45	-0.62	0.13
D 46	0.57	0.11
D 47	-1.19	0.14
D 48	0.13	0.11
D 51	-1.17	0.14
D 52	-1.38	0.15
E 53	0.23	0.11
E 54	2.29	0.10
E 56	-0.79	0.13
E 57	-0.99	0.14
E 58	-0.86	0.13

Content Word Items

ITEM NAME	ITEM DIFFIC.	STANDARD ERROR
I100	0.31	0.10
I104	1.33	0.11
I106	-0.06	0.10
I108	-1.37	0.12
J110	1.39	0.11
J111	4.03	0.19
J116	-0.78	0.11
J118	1.10	0.11
K122	2.03	0.12
K124	1.35	0.11
K125	0.06	0.10
K126	0.81	0.10
K127	-1.16	0.11
K128	1.41	0.11
K129	-0.14	0.10
L130	-1.78	0.12
L131	1.33	0.11
L133	3.03	0.14
L134	1.67	0.11
L135	-0.71	0.11
L136	2.33	0.12
L138	2.33	0.12
L141	-0.18	0.11

No. of persons = 600

Structure Word Items

ITEM NAME	ITEM DIFFIC.	STANDARD ERROR
E 59	2.87	0.11
E 60	-1.09	0.14
E 62	1.86	0.10
F 70	0.32	0.11
F 71	-0.57	0.13
F 72	-0.10	0.12
F 73	-0.83	0.13
F 76	1.92	0.10
G 77	1.75	0.10
G 80	-0.67	0.13
G 83	-0.84	0.13
G 84	-1.42	0.15
G 85	-1.11	0.14
G 87	0.45	0.11
H 88	-0.57	0.13
H 93	-0.81	0.13
I 99	-0.03	0.12
I101	3.79	0.12
I102	0.05	0.12
I103	0.70	0.11
I105	1.72	0.10
I107	-0.13	0.12
J109	4.88	0.16
J112	3.09	0.11
J113	0.10	0.12
J114	-1.31	0.15
J115	2.98	0.11
J117	1.25	0.10
J119	0.93	0.11
K120	0.17	0.11
K121	1.51	0.10
K123	2.49	0.10
L132	0.42	0.11
L137	0.98	0.11
L139	4.48	0.14
L140	-0.02	0.12

No. of persons = 601

G.3 Item Difficulty Estimates from Separate Calibrations of 'Open' & 'Closed' Items (Malaysian Data)

'Open' Items

ITEM NAME	ITEM DIFFIC.	STANDARD ERROR
A 3	-3.50	0.17
A 5	0.30	0.10
B 16	-0.05	0.10
B 18	-3.26	0.16
C 25	-4.71	0.24
C 29	-1.31	0.11
C 32	-2.74	0.14
C 35	-1.01	0.11
C 36	-1.06	0.11
D 44	-1.99	0.12
D 45	-2.20	0.13
E 53	-1.31	0.11
E 58	-2.42	0.13
E 59	1.31	0.11
E 62	0.30	0.10
E 63	-0.96	0.11
F 73	-2.37	0.13
G 79	0.41	0.10
G 80	-2.23	0.13
G 81	-2.40	0.13
G 82	1.32	0.11
G 87	-1.13	0.11
H 89	2.26	0.12
H 91	0.36	0.10
H 92	2.00	0.12
H 94	0.61	0.10
H 95	1.90	0.12
I 98	3.30	0.16
I104	1.20	0.11
I105	0.17	0.10
I106	-0.16	0.10
I108	-1.42	0.11
J110	1.26	0.11
J112	1.46	0.11
J115	1.39	0.11
J117	-0.34	0.10
J118	0.95	0.10
K122	1.86	0.11
K123	0.94	0.10
K124	1.21	0.11
K125	-0.05	0.10
K126	0.67	0.10
K128	1.28	0.11
K129	-0.21	0.10
L131	1.20	0.11
L133	2.82	0.14
L135	-0.78	0.11
L136	2.14	0.12
L138	2.16	0.12
L139	2.82	0.14

'Closed' Items

ITEM NAME	ITEM DIFFIC.	STANDARD ERROR
A 1	-2.93	0.21
A 2	-2.80	0.20
A 4	2.31	0.11
A 6	-0.51	0.12
A 7	1.04	0.10
A 8	-2.13	0.17
A 9	-2.57	0.19
A 10	-2.88	0.21
A 11	-2.07	0.17
B 12	-1.29	0.14
B 13	-0.51	0.12
B 14	0.35	0.11
B 15	-0.42	0.12
B 17	1.41	0.10
B 19	-1.21	0.14
B 20	1.29	0.10
B 21	-1.53	0.15
B 22	-1.69	0.15
B 23	-1.65	0.15
B 24	-0.52	0.12
C 26	-1.31	0.14
C 27	1.33	0.10
C 28	3.21	0.12
C 30	-1.31	0.14
C 31	-2.25	0.17
C 33	-2.97	0.21
C 34	-0.89	0.13
C 37	0.18	0.11
D 38	-1.65	0.15
D 39	-0.28	0.12
D 40	-0.85	0.13
D 41	-1.99	0.16
D 42	0.34	0.11
D 43	-0.42	0.12
D 46	0.37	0.11
D 47	-1.43	0.14
D 48	-0.06	0.11
D 49	0.43	0.11
D 50	1.21	0.10
D 51	-1.39	0.14
D 52	-1.56	0.15
E 54	2.12	0.10
E 55	-0.66	0.13
E 56	-0.97	0.13
E 57	-1.13	0.14
E 60	-1.27	0.14
E 61	0.83	0.11
E 64	2.03	0.10
F 65	-0.16	0.12
F 66	1.37	0.10
F 67	-0.08	0.12
F 68	1.15	0.10
F 69	1.19	0.10
F 70	0.15	0.11
F 71	-0.76	0.13
F 72	-0.28	0.12
F 74	0.73	0.11
F 75	0.45	0.11
F 76	1.74	0.10

'Open' Items

ITEM NAME	ITEM DIFFIC.	STANDARD ERROR

No. of persons = 598

'Closed' Items

ITEM NAME	ITEM DIFFIC.	STANDARD ERROR
G 77	1.59	0.10
G 78	2.31	0.11
G 83	-1.03	0.13
G 84	-1.60	0.15
G 85	-1.29	0.14
G 86	2.24	0.11
H 88	-0.74	0.13
H 90	0.36	0.11
H 93	-0.97	0.13
H 96	0.31	0.11
H 97	0.76	0.11
I 99	-0.18	0.12
I100	1.59	0.10
I101	3.72	0.12
I102	-0.13	0.12
I103	0.53	0.11
I107	-0.35	0.12
J109	4.81	0.16
J111	5.52	0.20
J113	-0.05	0.11
J114	-1.53	0.15
J116	0.50	0.11
J119	0.77	0.11
K120	0.03	0.11
K121	1.36	0.10
K127	0.12	0.11
L130	-0.46	0.12
L132	0.21	0.11
L134	3.00	0.11
L137	0.79	0.11
L140	-0.20	0.12
L141	1.17	0.10

No. of persons = 601

G.4 Item Difficulty Estimates for Score-Matched Malaysian & Tanzanian Groups

Score-Matched Malaysian Group

ITEM NAME	ITEM DIFFIC.	STANDARD ERROR
A 1	-3.58	0.24
A 2	-3.25	0.22
A 3	-2.52	0.19
A 4	2.14	0.22
A 5	1.31	0.18
A 6	-1.34	0.17
A 7	0.21	0.17
A 8	-2.68	0.20
A 9	-3.16	0.22
A 10	-3.25	0.22
A 11	-2.68	0.20
B 12	-1.57	0.17
B 13	-2.20	0.19
B 14	-0.17	0.16
B 15	-0.49	0.16
B 16	0.86	0.17
B 17	0.57	0.17
B 18	-2.72	0.20
B 19	-1.87	0.18
B 20	1.05	0.18
B 21	-1.87	0.18
B 22	-2.45	0.19
B 23	-1.78	0.18
B 24	-1.22	0.17
C 25	-4.04	0.28
C 26	-1.66	0.17
C 27	1.11	0.18
C 28	2.82	0.26
C 29	-0.30	0.16
C 30	-1.75	0.18
C 31	-3.11	0.22
C 32	-2.13	0.18
C 33	-3.46	0.24
C 34	-2.13	0.18
C 35	-0.14	0.16
C 36	-0.46	0.16
C 37	-0.41	0.16
D 38	-2.00	0.18
D 39	-0.84	0.16
D 40	-1.69	0.17
D 41	-2.49	0.19
D 42	-0.28	0.16
D 43	-0.73	0.16
D 44	-1.57	0.17
D 45	-1.00	0.17
D 46	0.18	0.17
D 47	-2.10	0.18
D 48	-0.76	0.16
D 49	-0.14	0.16
D 50	1.11	0.18

Score-Matched Tanzanian Group

ITEM NAME	ITEM DIFFIC.	STANDARD ERROR
A 1	-4.17	0.28
A 2	-2.65	0.20
A 3	-2.17	0.19
A 4	3.89	0.40
A 5	0.52	0.17
A 6	-1.21	0.17
A 7	1.00	0.18
A 8	-2.17	0.19
A 9	-3.11	0.22
A 10	-3.41	0.23
A 11	-3.57	0.24
B 12	-1.47	0.17
B 13	-2.49	0.19
B 14	-1.04	0.17
B 15	-0.57	0.17
B 16	0.91	0.18
B 17	0.09	0.17
B 18	-2.81	0.21
B 19	-1.81	0.18
B 20	2.36	0.23
B 21	-1.18	0.17
B 22	-2.21	0.19
B 23	-1.87	0.18
B 24	-1.33	0.17
C 25	-3.75	0.25
C 26	-1.50	0.17
C 27	-0.21	0.17
C 28	0.97	0.18
C 29	-0.87	0.17
C 30	-1.94	0.18
C 31	-2.89	0.21
C 32	-1.04	0.17
C 33	-1.75	0.18
C 34	-0.79	0.17
C 35	-0.30	0.17
C 36	-1.84	0.18
C 37	-0.87	0.17
D 38	-1.45	0.17
D 39	-1.15	0.17
D 40	-1.87	0.18
D 41	-1.78	0.18
D 42	0.63	0.17
D 43	-2.24	0.19
D 44	-0.46	0.17
D 45	-1.07	0.17
D 46	1.37	0.19
D 47	-1.27	0.17
D 48	-0.49	0.17
D 49	-1.07	0.17
D 50	1.07	0.18

Score-Matched Malaysian Group

ITEM NAME	ITEM DIFFIC.	STANDARD ERROR
D 51	-1.57	0.17
D 52	-2.30	0.19
E 53	-0.38	0.16
E 54	1.05	0.18
E 55	-1.51	0.17
E 56	-1.17	0.17
E 57	-1.60	0.17
E 58	-1.42	0.17
E 59	1.96	0.21
E 60	-1.87	0.18
E 61	0.32	0.17
E 62	1.52	0.19
E 63	-0.04	0.16
E 64	1.67	0.20
F 65	-1.34	0.17
F 66	0.95	0.18
F 67	-1.03	0.17
F 68	0.51	0.17
F 69	0.86	0.17
F 70	-0.54	0.16
F 71	-1.17	0.17
F 72	-0.86	0.17
F 73	-1.97	0.18
F 74	0.23	0.17
F 75	0.18	0.17
F 76	1.27	0.18
G 77	1.21	0.18
G 78	2.44	0.24
G 79	1.48	0.19
G 80	-1.28	0.17
G 81	-1.72	0.18
G 82	1.83	0.20
G 83	-1.28	0.17
G 84	-2.07	0.18
G 85	-1.72	0.18
G 86	2.18	0.22
G 87	-0.36	0.16
H 88	-1.14	0.17
H 89	3.04	0.28
H 90	-0.09	0.16
H 91	1.67	0.20
H 92	2.89	0.27
H 93	-1.17	0.17
H 94	1.34	0.19
H 95	2.44	0.24
H 96	-0.04	0.16
H 97	-0.04	0.16
I 98	3.64	0.35
I 99	-0.70	0.16

Score-Matched Tanzanian Group

ITEM NAME	ITEM DIFFIC.	STANDARD ERROR
D 51	-1.36	0.17
D 52	-1.36	0.17
E 53	0.34	0.17
E 54	1.30	0.19
E 55	-1.53	0.17
E 56	-2.10	0.18
E 57	-0.85	0.17
E 58	-1.01	0.17
E 59	-2.69	0.20
E 60	-1.66	0.18
E 61	0.66	0.17
E 62	1.63	0.20
E 63	-0.87	0.17
E 64	2.25	0.23
F 65	-2.38	0.19
F 66	-0.41	0.17
F 67	-0.65	0.17
F 68	-0.60	0.17
F 69	-0.02	0.17
F 70	-1.21	0.17
F 71	-0.30	0,17
F 72	-1.18	0.17
F 73	-2.28	0.19
F 74	0.82	0.18
F 75	0.32	0.17
F 76	0.01	0.17
G 77	1.56	0.19
G 78	1.60	0.20
G 79	1.48	0.19
G 80	-1.56	0.17
G 81	-2.93	0.21
G 82	1.52	0.19
G 83	-2.28	0.19
G 84	-1.21	0.17
G 85	-1.87	0.18
G 86	0.63	0.17
G 87	0.26	0.17
H 88	-2.04	0.18
H 89	1.30	0.19
H 90	-0.93	0.17
H 91	1.79	0.20
H 92	3.89	0.40
H 93	-1.69	0.18
H 94	1.84	0.21
H 95	3.01	0.28
H 96	0.91	0.18
H 97	1.67	0.20
I 98	5.93	1.00
I 99	-0.46	0.17

Score-Matched Malaysian Group

ITEM NAME	ITEM DIFFIC.	STANDARD ERROR
I100	1.56	0.19
I101	3.12	0.29
I102	-0.44	0.16
I103	0.40	0.17
I104	2.34	0.23
I105	1.27	0.18
I106	0.95	0.18
I107	-0.46	0.16
I108	-0.28	0.16
J109	3.77	0.37
J110	2.44	0.24
J111	4.52	0.51
J112	1.87	0.20
J113	-0.62	0.16
J114	-1.81	0.18
J115	2.23	0.22
J116	0.32	0.17
J117	0.45	0.17
J118	2.28	0.23
J119	0.43	0.17
K120	-0.62	0.16
K121	0.57	0.17
K122	3.12	0.29
K123	1.38	0.19
K124	2.68	0.25
K125	0.98	0.18
K126	1.79	0.20
K127	-0.17	0.16
K128	2.28	0.23
K129	0.80	0.17
L130	-0.76	0.16
L131	2.56	0.24
L132	-0.06	0.16
L133	4.09	0.42
L134	2.89	0.27
L135	0.34	0.17
L136	2.44	0.24
L137	0.71	0.17
L138	3.52	0.34
L139	3.64	0.35
L140	-0.76	0.16
L141	0.51	0.17

No. of persons = 230

Score-Matched Tanzanian Group

ITEM NAME	ITEM DIFFIC.	STANDARD ERROR
I100	1.79	0.20
I101	2.65	0.25
I102	-0.32	0.17
I103	-0.62	0.17
I104	0.79	0.18
I105	2.25	0.23
I106	1.84	0.21
I107	0.23	0.17
I108	0.09	0.17
J109	5.93	1.00
J110	1.92	0.21
J111	5.22	0.71
J112	1.97	0.21
J113	-0.71	0.17
J114	-1.91	0.18
J115	1.48	0.19
J116	0.03	0.17
J117	1.17	0.18
J118	1.71	0.20
J119	0.29	0.17
K120	0.20	0.17
K121	0.85	0.18
K122	3.75	0.37
K123	1.41	0.19
K124	1.88	0.21
K125	1.37	0.19
K126	0.82	0.18
K127	0.52	0.17
K128	1.97	0.21
K129	1.13	0.18
L130	3.75	0.37
L131	1.63	0.20
L132	-0.51	0.17
L133	2.41	0.24
L134	2.10	0.22
L135	0.43	0.17
L136	2.72	0.26
L137	0.94	0.18
L138	3.19	0.30
L139	3.01	0.28
L140	-0.38	0.17
L141	0.72	0.17

No. of persons = 227

G.5 Item Difficulty Estimates from Separate Calibrations of Hard & Easy Item Subsets (Malaysian Data)

40 Hardest Items

ITEM NAME	ITEM DIFFIC.	STANDARD ERROR
A 4	-0.50	0.11
A 5	-1.17	0.11
C 28	0.39	0.12
E 54	-0.73	0.11
E 59	-0.10	0.11
E 62	-1.14	0.11
E 64	-0.81	0.11
F 76	-1.11	0.11
G 77	-1.27	0.11
G 78	-0.50	0.11
G 79	-1.03	0.11
G 82	-0.06	0.11
G 86	-0.57	0.11
H 89	0.95	0.13
H 91	-1.07	0.11
H 92	0.68	0.12
H 94	-0.79	0.11
H 95	0.54	0.12
I 98	2.08	0.17
I100	-1.28	0.11
I101	0.89	0.12
I104	-0.18	0.11
I105	-1.30	0.11
J109	1.98	0.16
J110	-0.10	0.11
J111	2.59	0.20
J112	0.11	0.11
J115	0.02	0.11
J118	-0.42	0.11
K122	0.54	0.12
K123	-0.49	0.11
K124	-0.17	0.11
K126	-0.71	0.11
K128	-0.10	0.11
L131	-0.16	0.11
L133	1.56	0.14
L134	0.19	0.11
L136	0.83	0.12
L138	0.85	0.12
L139	1.54	0.14

No. of persons = 547

40 Easiest Items

ITEM NAME	ITEM DIFFIC.	STANDARD ERROR
A 1	-1.35	0.21
A 2	-1.27	0.21
A 3	-0.52	0.17
A 8	-0.54	0.17
A 9	-1.03	0.19
A 10	-1.35	0.21
A 11	-0.52	0.17
B 12	0.31	0.14
B 18	-0.27	0.16
B 19	0.41	0.14
B 21	0.08	0.15
B 22	-0.08	0.16
B 23	-0.01	0.15
C 25	-1.79	0.24
C 26	0.31	0.14
C 30	0.33	0.14
C 31	-0.66	0.18
C 32	0.29	0.14
C 33	-1.44	0.22
C 34	0.73	0.13
D 38	-0.01	0.15
D 40	0.77	0.13
D 41	-0.40	0.17
D 45	0.83	0.13
D 47	0.23	0.15
D 51	0.25	0.14
D 52	0.04	0.15
E 56	0.63	0.13
E 57	0.47	0.14
E 58	0.56	0.14
E 60	0.35	0.14
F 71	0.88	0.13
F 73	0.63	0.13
G 80	0.77	0.13
G 81	0.65	0.13
G 83	0.61	0.14
G 84	0.04	0.15
G 85	0.33	0.14
H 93	0.63	0.13
J114	0.10	0.15

No. of persons = 481

APPENDIX H
TRADITIONAL STATISTICS FOR ELTS TEST

H.1 ELTS Subtests: Raw Score Distributions & Frequency Counts, K-R20 & SEM

Table 1: G1 (Reading)

RAW SCORE	FREQ.	CUM. FREQ.	NO. OF PERSONS (0 | 20 | 40 | 60 | 80 | 100)
0	0	0	|
1	0	0	|
2	0	0	|
3	0	0	|
4	0	0	|
5	0	0	|
6	0	0	|
7	3	3	|**
8	3	6	|**
9	2	8	|*
10	4	12	|**
11	3	15	|**
12	8	23	|****
13	8	31	|****
14	13	44	|*******
15	16	60	|********
16	32	92	|******************
17	31	123	|******************
18	26	149	|**************
19	39	188	|**********************
20	31	219	|******************
21	42	261	|************************
22	61	322	|**********************************
23	51	373	|*****************************
24	72	445	|**
25	72	517	|**
26	74	591	|***
27	76	667	|**
28	88	755	|**
29	62	817	|**********************************
30	86	903	|**
31	58	961	|********************************
32	71	1032	|**
33	56	1088	|********************************
34	74	1162	|***
35	75	1237	|**
36	84	1321	|**
37	61	1382	|**********************************
38	58	1440	|********************************
39	40	1480	|**********************
40	23	1503	|************

Mean raw score = 28.21
SD raw scores = 6.94
Raw score range for group: 7 to 40

K-R20 = 0.87
SEM = 2.53

Table 2: G2 (Listening)

RAW SCORE	FREQ.	CUM. FREQ.	NO. OF PERSONS (0 20 40 60 80 100 120)
0	0	0	|
1	0	0	|
2	0	0	|
3	0	0	|
4	1	1	|*
5	1	2	|*
6	0	2	|
7	1	3	|*
8	3	6	|**
9	5	11	|***
10	6	17	|***
11	12	29	|******
12	20	49	|**********
13	18	67	|*********
14	20	87	|**********
15	29	116	|**********
16	45	161	|**********
17	51	212	|**********
18	63	275	|**********
19	73	348	|**********
20	63	411	|**********
21	95	506	|**********
22	96	602	|**********
23	99	701	|**********
24	104	805	|**********
25	105	910	|**********
26	118	1028	|**********
27	107	1135	|**********
28	78	1213	|**********
29	77	1290	|**********
30	83	1373	|**********
31	41	1414	|**********
32	41	1455	|**********
33	37	1492	|**********
34	11	1503	|******
35	0	1503	|

Mean raw score = 23.54
SD raw scores = 5.38
Raw score range: 4 to 34

K-R20 = 0.80
SEM = 2.42

Table 3: M1 (General Academic)

RAW SCORE	FREQ.	CUM. FREQ.	NO. OF PERSONS (0 | 10 | 20 | 30)
0	0	0	|
1	0	0	|
2	0	0	|
3	0	0	|
4	0	0	|
5	4	4	|****
6	3	7	|***
7	5	12	|*****
8	3	15	|***
9	5	20	|*****
10	10	30	|**********
11	14	44	|**************
12	19	63	|*******************
13	18	81	|******************
14	14	95	|**************
15	18	113	|******************
16	21	134	|*********************
17	16	150	|****************
18	19	169	|*******************
19	16	185	|****************
20	23	208	|***********************
21	17	225	|*****************
22	17	242	|*****************
23	16	258	|****************
24	8	266	|********
25	12	278	|************
26	13	291	|*************
27	12	303	|************
28	16	319	|****************
29	9	328	|*********
30	10	338	|**********
31	7	345	|*******
32	12	357	|************
33	8	365	|********
34	8	373	|********
35	6	379	|******
36	12	391	|************
37	6	397	|******
38	3	400	|***
39	2	402	|**
40	1	403	|*

Mean raw score = 21.17
SD raw scores = 8.15
Raw score range: 5 to 40

K-R20 = 0.88
SEM = 2.78

Table 4 : M1 (Life Sciences)

RAW SCORE	FREQ.	CUM. FREQ.	NO. OF PERSONS 0 10 20 30
0	1	1	*
1	0	1	
2	0	1	
3	0	1	
4	0	1	
5	0	1	
6	0	1	
7	0	1	
8	1	2	*
9	1	3	*
10	2	5	**
11	2	7	**
12	0	7	
13	3	10	***
14	6	16	******
15	8	24	********
16	12	36	************
17	5	41	*****
18	16	57	*****************
19	25	82	*************************
20	26	108	**************************
21	18	126	******************
22	15	141	***************
23	26	167	**************************
24	17	184	*****************
25	23	207	************************
26	11	218	**********
27	26	244	**************************
28	30	274	**********************************
29	15	289	***************
30	17	306	*****************
31	13	319	************
32	14	333	**************
33	14	347	**************
34	7	354	*******
35	7	361	*******
36	10	371	**********
37	3	374	***
38	0	374	
39	0	374	
40	0	374	

Mean raw score = 24.57
SD raw scores = 6.14
Raw score range: 0 to 37

K-R20 = 0.81
SEM = 2.68

Table 5: M1 (Medicine)

RAW SCORE	FREQ.	CUM. FREQ.	NO. OF PERSONS 0 10 20 30
0	0	0	|
1	0	0	|
2	0	0	|
3	0	0	|
4	0	0	|
5	0	0	|
6	0	0	|
7	0	0	|
8	1	1	|*
9	3	4	|***
10	0	4	|
11	1	5	|*
12	0	5	|
13	0	5	|
14	1	6	|*
15	0	6	|
16	2	8	|**
17	1	9	|*
18	0	9	|
19	2	11	|**
20	4	15	|****
21	4	19	|****
22	6	25	|******
23	5	30	|*****
24	4	34	|****
25	5	39	|*****
26	7	46	|*******
27	8	54	|********
28	6	60	|******
29	9	69	|*********
30	9	78	|*********
31	9	87	|*********
32	8	95	|********
33	12	107	|************
34	15	122	|***************
35	6	128	|******
36	6	134	|******
37	6	140	|******
38	3	143	|***
39	0	143	|
40	0	143	|

Mean raw score = 28.52
SD raw scores = 6.48
Raw score range: 8 to 38

K-R20 = 0.85
SEM = 2.48

Table 6: M1 (Physical Sciences)

RAW SCORE	FREQ.	CUM. FREQ.	NO. OF PERSONS 0 10 20
0	1	1	*
1	0	1	
2	0	1	
3	0	1	
4	0	1	
5	0	1	
6	0	1	
7	0	1	
8	2	3	**
9	0	3	
10	1	4	*
11	0	4	
12	0	4	
13	0	4	
14	0	4	
15	1	5	*
16	1	6	*
17	3	9	***
18	0	9	
19	4	13	****
20	1	14	*
21	6	20	******
22	1	21	*
23	5	26	*****
24	0	26	
25	7	33	*******
26	2	35	**
27	10	45	***********
28	2	47	**
29	4	51	****
30	4	55	****
31	8	63	********
32	11	74	************
33	8	82	********
34	11	93	************
35	9	102	*********
36	9	111	*********
37	12	123	*************
38	8	131	********
39	3	134	***
40	0	134	

Mean raw score = 29.84
SD raw scores = 7.26
Raw score range: 0 to 39

K-R20 = 0.90
SEM = 2.31

Table 7: M1 (Social Studies)

RAW SCORE	FREQ.	CUM. FREQ.	NO. OF PERSONS 0 10 20
0	0	0	|
1	0	0	|
2	0	0	|
3	0	0	|
4	0	0	|
5	0	0	|
6	3	3	***
7	0	3	|
8	2	5	**
9	1	6	*
10	5	11	*****
11	5	16	*****
12	7	23	*******
13	4	27	****
14	5	32	*****
15	20	52	********************
16	7	59	*******
17	9	68	*********
18	10	78	**********
19	16	94	*****************
20	16	110	*****************
21	15	125	****************
22	17	142	******************
23	13	155	*************
24	18	173	******************
25	15	188	****************
26	15	203	****************
27	9	212	*********
28	6	218	******
29	9	227	*********
30	8	235	********
31	9	244	*********
32	2	246	**
33	4	250	****
34	7	257	*******
35	3	260	***
36	1	261	*
37	1	262	*
38	1	263	*
39	1	264	*
40	0	264	|

Mean raw score = 21.92
SD raw scores = 6.62
Raw score range: 6 to 39

K-R20 = 0.82
SEM = 2.81

Table 8: M1 (Technology)

RAW SCORE	FREQ.	CUM. FREQ.	NO. OF PERSONS 0 10 20
0	0	0	|
1	0	0	|
2	0	0	|
3	0	0	|
4	0	0	|
5	0	0	|
6	4	4	|****
7	2	6	|**
8	1	7	|*
9	0	7	|
10	1	8	|*
11	0	8	|
12	0	8	|
13	0	8	|
14	1	9	|*
15	1	10	|*
16	0	10	|
17	1	11	|*
18	2	13	|**
19	3	16	|***
20	0	16	|
21	3	19	|***
22	5	24	|*****
23	0	24	|
24	4	28	|****
25	7	35	|*******
26	7	42	|*******
27	3	45	|***
28	11	56	|**********
29	14	70	|**************
30	13	83	|*************
31	10	93	|**********
32	15	108	|***************
33	20	128	|********************
34	13	141	|*************
35	15	156	|***************
36	12	168	|************
37	8	176	|********
38	4	180	|****
39	4	184	|****
40	1	185	|*

Mean raw score = 29.72
SD raw scores = 6.96
Raw score range: 6 to 40

K-R20 = 0.88
SEM = 2.43

H.2 ELTS Subtests: Traditional Item Statistics

Table 1: G1 (Reading)

ITEM LABEL	FACILITY VALUE	DISCRIM INDEX	UNBIASED PT.BISERIAL
G101	0.93	0.20	0.30
G102	0.91	0.21	0.30
G103	0.80	0.32	0.28
G104	0.88	0.27	0.30
G105	0.83	0.29	0.26
G106	0.86	0.21	0.20
G107	0.82	0.25	0.21
G108	0.86	0.25	0.26
G109	0.59	0.40	0.25
G110	0.61	0.61	0.41
G111	0.41	0.31	0.17
G112	0.49	0.44	0.30
G113	0.87	0.32	0.36
G114	0.76	0.49	0.39
G115	0.74	0.44	0.36
G116	0.76	0.38	0.32
G117	0.59	0.55	0.41
G118	0.68	0.50	0.35
G119	0.90	0.19	0.24
G120	0.88	0.25	0.28
G121	0.77	0.45	0.40
G122	0.83	0.35	0.35
G123	0.77	0.48	0.41
G124	0.66	0.47	0.34
G125	0.43	0.38	0.25
G126	0.44	0.58	0.39
G127	0.86	0.31	0.35
G128	0.69	0.39	0.30
G129	0.76	0.44	0.37
G130	0.70	0.48	0.35
G131	0.62	0.52	0.37
G132	0.58	0.51	0.34
G133	0.80	0.48	0.46
G134	0.73	0.55	0.46
G135	0.50	0.66	0.46
G136	0.71	0.57	0.45
G137	0.54	0.77	0.54
G138	0.40	0.71	0.50
G139	0.64	0.55	0.41
G140	0.60	0.67	0.50

No. of persons = 1,503

Table 2: G2 (Listening)

ITEM LABEL	FACILITY VALUE	DISCRIM INDEX	UNBIASED PT.BISERIAL
G201	0.93	0.15	0.22
G202	0.71	0.42	0.31
G203	0.59	0.42	0.25
G204	0.56	0.51	0.35
G205	0.83	0.33	0.29
G206	0.73	0.48	0.36
G207	0.70	0.43	0.30
G208	0.72	0.45	0.33
G209	0.64	0.52	0.35
G210	0.73	0.42	0.33
G211	0.67	0.42	0.28
G212	0.80	0.40	0.35
G213	0.41	0.47	0.29
G214	0.55	0.55	0.34
G215	0.80	0.35	0.31
G216	0.54	0.56	0.36
G217	0.86	0.27	0.26
G218	0.90	0.27	0.35
G219	0.86	0.27	0.26
G220	0.94	0.15	0.24
G221	0.81	0.34	0.31
G222	0.69	0.52	0.39
G223	0.75	0.46	0.36
G224	0.83	0.33	0.32
G225	0.54	0.49	0.29
G226	0.83	0.22	0.18
G227	0.10	-0.06	-0.15
G228	0.57	0.48	0.31
G229	0.61	0.47	0.30
G230	0.30	0.42	0.29
G231	0.84	0.27	0.24
G232	0.67	0.42	0.27
G233	0.59	0.40	0.22
G234	0.61	0.44	0.28
G235	0.30	0.22	0.13

No. of persons = 1,503

Table 3: M1 (General Academic)

ITEM LABEL	FACILITY VALUE	DISCRIM INDEX	UNBIASED PT.BISERIAL
GA01	0.76	0.23	0.19
GA02	0.48	0.51	0.31
GA03	0.37	-0.03	-0.08
GA04	0.62	0.53	0.39
GA05	0.51	0.19	0.12
GA06	0.66	0.43	0.29
GA07	0.75	0.46	0.38
GA08	0.73	0.35	0.26
GA09	0.42	0.56	0.40
GA10	0.68	0.49	0.33
GA11	0.84	0.22	0.24
GA12	0.75	0.34	0.26
GA13	0.45	0.39	0.29
GA14	0.55	0.61	0.43
GA15	0.59	0.47	0.31
GA16	0.52	0.39	0.30
GA17	0.63	0.29	0.20
GA18	0.37	0.52	0.41
GA19	0.51	0.47	0.32
GA20	0.61	0.72	0.52
GA21	0.38	0.50	0.41
GA22	0.75	0.47	0.38
GA23	0.62	0.61	0.46
GA24	0.57	0.77	0.56
GA25	0.37	0.59	0.42
GA26	0.29	0.36	0.26
GA27	0.57	0.50	0.33
GA28	0.66	0.56	0.44
GA29	0.48	0.60	0.42
GA30	0.37	0.43	0.33
GA31	0.55	0.62	0.44
GA32	0.57	0.62	0.44
GA33	0.38	0.64	0.49
GA34	0.48	0.79	0.59
GA35	0.31	0.49	0.41
GA36	0.50	0.79	0.55
GA37	0.40	0.82	0.61
GA38	0.42	0.65	0.47
GA39	0.30	0.63	0.55
GA40	0.40	0.79	0.59

No. of persons = 403

Table 4: M1 (Life Sciences)

ITEM LABEL	FACILITY VALUE	DISCRIM INDEX	UNBIASED PT.BISERIAL
LS01	0.81	0.20	0.14
LS02	0.89	0.18	0.17
LS03	0.71	0.29	0.23
LS04	0.50	0.04	0.02
LS05	0.95	0.10	0.21
LS06	0.93	0.08	0.17
LS07	0.80	0.35	0.28
LS08	0.33	0.14	0.05
LS09	0.83	0.25	0.18
LS10	0.65	0.34	0.21
LS11	0.72	0.20	0.14
LS12	0.79	0.28	0.21
LS13	0.82	0.28	0.26
LS14	0.45	0.32	0.17
LS15	0.80	0.23	0.24
LS16	0.56	0.24	0.10
LS17	0.59	0.40	0.28
LS18	0.80	0.25	0.19
LS19	0.72	0.25	0.21
LS20	0.67	0.35	0.25
LS21	0.73	0.36	0,25
LS22	0.43	0.35	0.19
LS23	0.86	0.29	0.30
LS24	0.86	0.31	0.34
LS25	0.69	0.42	0.28
LS26	0.53	0.50	0.31
LS27	0.47	0.64	0.43
LS28	0.45	0.51	0.37
LS29	0.66	0.65	0.48
LS30	0.56	0.58	0.40
LS31	0.51	0.67	0.50
LS32	0.43	0.62	0.45
LS33	0.53	0.63	0.44
LS34	0.37	0.54	0.36
LS35	0.05	0.01	-0.01
LS36	0.38	0.45	0.29
LS37	0.44	0.71	0.47
LS38	0.38	0.63	0.44
LS39	0.50	0.75	0.50
LS40	0.46	0.67	0.47

No of persons = 374

Table 5: M1 (Medicine)

ITEM LABEL	FACILITY VALUE	DISCRIM INDEX	UNBIASED PT.BISERIAL
ME01	0.89	0.28	0.34
ME02	0.92	0.23	0.41
ME03	0.81	0.38	0.45
ME04	0.77	0.31	0.19
ME05	0.69	0.23	0.15
ME06	0.36	0.33	0.22
ME07	0.85	0.18	0.27
ME08	0.87	0.21	0.24
ME09	0.86	0.28	0.42
ME10	0.81	0.36	0.43
ME11	0.92	0.13	0.25
ME12	0.92	0.18	0.29
ME13	0.30	0.26	0.17
ME14	0.81	0.28	0.25
ME15	0.77	0.36	0.30
ME16	0.85	0.28	0.33
ME17	0.94	0.21	0.46
ME18	0.92	0.15	0.26
ME19	0.84	0.41	0.40
ME20	0.83	0.23	0.08
ME21	0.48	0.54	0.35
ME22	0.85	0.15	0.17
ME23	0.66	0.69	0.47
ME24	0.55	0.51	0.35
ME25	0.57	0.46	0.27
ME26	0.54	0.46	0.27
ME27	0.62	0.74	0.45
ME28	0.53	0.54	0.32
ME29	0.76	0.51	0.45
ME30	0.55	0.59	0.41
ME31	0.82	0.46	0.53
ME32	0.57	0.46	0.30
ME33	0.90	0.28	0.48
ME34	0.70	0.56	0.54
ME35	0.38	0.28	0.15
ME36	0.61	0.64	0.45
ME37	0.73	0.44	0.39
ME38	0.73	0.46	0.44
ME39	0.51	0.62	0.44
ME40	0.55	0.49	0.36

No. of persons = 143

Table 6: M1 (Physical Sciences)

ITEM LABEL	FACILITY VALUE	DISCRIM INDEX	UNBIASED PT.BISERIAL
PS01	0.93	0.11	0.31
PS02	0.72	0.58	0.50
PS03	0.87	0.42	0.49
PS04	0.37	0.58	0.37
PS05	0.81	0.42	0.46
PS06	0.60	0.31	0.27
PS07	0.49	0.61	0.41
PS08	0.90	0.22	0.37
PS09	0.96	0.17	0.54
PS10	0.79	0.47	0.48
PS11	0.93	0.17	0.43
PS12	0.69	0.39	0.35
PS13	0.57	0.78	0.54
PS14	0.73	0.50	0.38
PS15	0.73	0.33	0.30
PS16	0.92	0.25	0.51
PS17	0.84	0.28	0.34
PS18	0.81	0.42	0.30
PS19	0.91	0.28	0.45
PS20	0.87	0.28	0.38
PS21	0.77	0.56	0.53
PS22	0.90	0.28	0.34
PS23	0.85	0.36	0.46
PS24	0.82	0.28	0.31
PS25	0.81	0.36	0.31
PS26	0.77	0.56	0.51
PS27	0.88	0.31	0.43
PS28	0.78	0.42	0.42
PS29	0.63	0.67	0.52
PS30	0.87	0.33	0.45
PS31	0.62	0.67	0.48
PS32	0.67	0.44	0.30
PS33	0.73	0.61	0.48
PS34	0.86	0.44	0.53
PS35	0.55	0.53	0.35
PS36	0.75	0.58	0.46
PS37	0.71	0.64	0.51
PS38	0.69	0.72	0.53
PS39	0.60	0.56	0.35
PS40	0.15	0.03	0.05

No. of persons = 134

Table 7: M1 (Social Studies)

ITEM LABEL	FACILITY VALUE	DISCRIM INDEX	UNBIASED PT.BISERIAL
SS01	0.47	0.52	0.37
SS02	0.61	0.48	0.31
SS03	0.58	0.46	0.31
SS04	0.69	0.20	0.10
SS05	0.31	0.49	0.33
SS06	0.61	0.25	0.21
SS07	0.66	0.45	0.31
SS08	0.38	0.42	0.28
SS09	0.80	0.41	0.32
SS10	0.58	0.37	0.27
SS11	0.55	0.61	0.44
SS12	0.39	0.52	0.36
SS13	0.68	0.44	0.36
SS14	0.57	0.21	0.10
SS15	0.80	0.28	0.21
SS16	0.91	0.18	0.21
SS17	0.92	0.17	0.21
SS18	0.88	0.31	0.36
SS19	0.16	-0.01	-0.04
SS20	0.65	0.34	0.22
SS21	0.45	0.42	0.28
SS22	0.41	0.21	0.12
SS23	0.52	0.52	0.35
SS24	0.50	0.28	0.16
SS25	0.57	0.39	0.23
SS26	0.56	0.38	0.24
SS27	0.70	0.51	0.37
SS28	0.62	0.62	0.43
SS29	0.38	0.51	0.35
SS30	0.44	0.61	0.41
SS31	0.53	0.62	0.41
SS32	0.47	0.65	0.45
SS33	0.59	0.65	0.48
SS34	0.41	0.42	0.30
SS35	0.55	0.49	0.34
SS36	0.31	0.21	0.12
SS37	0.61	0.58	0.40
SS38	0.38	0.54	0.37
SS39	0.45	0.46	0.31
SS40	0.28	0.27	0.17

No. of persons = 264

Table 8: M1 (Technology)

ITEM LABEL	FACILITY VALUE	DISCRIM INDEX	UNBIASED PT.BISERIAL
TN01	0.86	0.20	0.33
TN02	0.80	0.46	0.45
TN03	0.93	0.24	0.61
TN04	0.91	0.30	0.59
TN05	0.78	0.42	0.40
TN06	0.94	0.14	0.24
TN07	0.91	0.28	0.57
TN08	0.92	0.22	0.59
TN09	0.86	0.26	0.37
TN10	0.71	0.56	0.42
TN11	0.64	0.38	0.32
TN12	0.62	0.48	0.37
TN13	0.48	0.16	-0.10
TN14	0.84	0.38	0.46
TN15	0.89	0.34	0.63
TN16	0.83	0.44	0.51
TN17	0.68	0.50	0.44
TN18	0.72	0.50	0.42
TN19	0.75	0.42	0.34
TN20	0.77	0.44	0.44
TN21	0.55	0.36	0.31
TN22	0.62	0.38	0.17
TN23	0.66	0.62	0.46
TN24	0.70	0.20	0.17
TN25	0.86	0.38	0.53
TN26	0.78	0.36	0.45
TN27	0.68	0.22	0.07
TN28	0.68	0.40	0.14
TN29	0.80	0.32	0.41
TN30	0.75	0.28	0.33
TN31	0.79	0.54	0.55
TN32	0.92	0.26	0.62
TN33	0.61	0.60	0.41
TN34	0.85	0.42	0.59
TN35	0.55	0.54	0.39
TN36	0.74	0.54	0.54
TN37	0.75	0.50	0.34
TN38	0.60	0.24	0.22
TN39	0.42	0.40	0.18
TN40	0.60	0.58	0.37

No. of persons = 185

H.3 ELTS Subtests: Grouped Item Statistics

Table 1(a) ELTS Test (G1-Reading): Items Grouped by Facility Value

Facility Value Interval	No.of Items	Item Names
0.90-1.00	3	G101 G102 G119
0.80-0.89	11	G103 G104 G105 G106 G107 G108 G113 G120 G122 G127 G133
0.70-0.79	9	G114 G115 G116 G121 G123 G129 G130 G134 G136
0.60-0.69	7	G110 G118 G124 G128 G131 G139 G140
0.50-0.59	5	G109 G117 G132 G135 G137
0.40-0.49	5	G111 G112 G125 G126 G138
0.30-0.39	0	
0.20-0.29	0	
0.10-0.19	0	
0.00-0.09	0	

Facility value range = 0.40 (Item G138) to 0.93 (Item G101)
Mean = 0.71
SD = 0.15

Table 1(b) ELTS Test (G1-Reading): Items Grouped by Discrimination Index

Discrimination Index Interval	No.of Items	Item Names
0.90-1.00	0	
0.80-0.89	0	
0.70-0.79	2	G137 G138
0.60-0.69	3	G110 G135 G140
0.50-0.59	8	G117 G118 G126 G131 G132 G134 G136 G139
0.40-0.49	10	G109 G112 G114 G115 G121 G123 G124 G129 G130 G133
0.30-0.39	8	G103 G111 G113 G116 G122 G125 G127 G128
0.20-0.29	8	G101 G102 G104 G105 G106 G107 G108 G120
0.10-0.19	1	G119
0.00-0.09	0	

Discrimination index range = 0.19 (Item G119) to 0.77 (Item G137)
Mean = 0.43
SD = 0.15

Table 1(c) ELTS Test (G1-Reading): Items Grouped by Unbiased Point Biserial

Point Biserial Interval	No.of Items	Item Names
0.90-1.00	0	
0.80-0.89	0	
0.70-0.79	0	
0.60-0.69	0	
0.50-0.59	3	G137 G138 G140
0.40-0.49	9	G110 G117 G121 G123 G133 G134 G135 G136 G139
0.30-0.39	18	G101 G102 G104 G112 G113 G114 G115 G116 G118 G122 G124 G126 G127 G128 G129 G130 G131 G132
0.20-0.29	9	G103 G105 G106 G107 G108 G109 G119 G120 G125
0.10-0.19	1	G111
0.00-0.09	0	

Point biserial range = 0.17 (Item G111) to 0.54 (Item G137)
Mean = 0.35
SD = 0.09

Table 2(a) ELTS Test (G2-Listening): Items Grouped by Facility Value

Facility Value Interval	No.of Items	Item Names
0.90-1.00	3	G201 G218 G220
0.80-0.89	9	G205 G212 G215 G217 G219 G221 G224 G226 G231
0.70-0.79	6	G202 G206 G207 G208 G210 G223
0.60-0.69	6	G209 G211 G222 G229 G232 G234
0.50-0.59	7	G203 G204 G214 G216 G225 G228 G233
0.40-0.49	1	G213
0.30-0.39	2	G230 G235
0.20-0.29	0	
0.10-0.19	1	G227
0.00-0.09	0	

Facility value range = 0.10 (Item G227) to 0.94 (Item G220)
Mean = 0.67
SD = 0.19

Table 2(b) ELTS Test (G2-Listening): Items Grouped by Discrimination Index

Discrimination Index Interval	No.of Items	Item Names
0.90-1.00	0	
0.80-0.89	0	
0.70-0.79	0	
0.60-0.69	0	
0.50-0.59	5	G204 G209 G214 G216 G222
0.40-0.49	17	G202 G203 G206 G207 G208 G210 G211 G212 G213 G223 G225 G228 G229 G230 G232 G233 G234
0.30-0.39	4	G205 G215 G221 G224
0.20-0.29	6	G217 G218 G219 G226 G231 G235
0.10-0.19	2	G201 G220
0.00-0.09	0	
-0.10- -0.01	1	G227

Discrimination index range = -0.06 (Item G227) to 0.56 (Item G216)
Mean = 0.38
SD = 0.13

Table 2(c) ELTS Test (G2-Listening): Items Grouped by Unbiased Point Biserial

Point Biserial Interval	No.of Items	Item Names
0.90-1.00	0	
0.80-0.89	0	
0.70-0.79	0	
0.60-0.69	0	
0.50-0.59	0	
0.40-0.49	0	
0.30-0.39	18	G202 G204 G206 G207 G208 G209 G210 G212 G214 G215 G216 G218 G221 G222 G223 G224 G228 G229
0.20-0.29	14	G201 G203 G205 G211 G213 G217 G219 G220 G225 G230 G231 G232 G233 G234
0.10-0.19	2	G226 G235
0.00-0.09	0	
-0.10- -0.01	0	
-0.20- -0.11	1	G227

Point biserial range = -0.15 (Item G227) to 0.39 (Item G222)
Mean = 0.28
SD = 0.09

Table 3(a) ELTS Test (M1-General Academic): Items Grouped by Facility Value

Facility Value Interval	No.of Items	Item Names
0.90-1.00	0	
0.80-0.89	1	GA11
0.70-0.79	5	GA01 GA07 GA08 GA12 GA22
0.60-0.69	7	GA04 GA06 GA10 GA17 GA20 GA23 GA28
0.50-0.59	10	GA05 GA14 GA15 GA16 GA19 GA24 GA27 GA31 GA32 GA36
0.40-0.49	8	GA02 GA09 GA13 GA29 GA34 GA37 GA38 GA40
0.30-0.39	8	GA03 GA18 GA21 GA25 GA30 GA33 GA35 GA39
0.20-0.29	1	GA26
0.10-0.19	0	
0.00-0.09	0	

Facility value range = 0.30 (Item GA39) to 0.84 (Item GA11)
Mean = 0.53
SD = 0.14

Table 3(b) ELTS Test (M1-General Academic): Items Grouped by Discrimination Index

Discrimination Index Interval	No.of Items	Item Names
0.90-1.00	0	
0.80-0.89	1	GA37
0.70-0.79	5	GA20 GA24 GA34 GA36 GA40
0.60-0.69	8	GA14 GA23 GA29 GA31 GA32 GA33 GA38 GA39
0.50-0.59	8	GA02 GA04 GA09 GA18 GA21 GA25 GA27 GA28
0.40-0.49	8	GA06 GA07 GA10 GA15 GA19 GA22 GA30 GA35
0.30-0.39	5	GA08 GA12 GA13 GA16 GA26
0.20-0.29	3	GA01 GA11 GA17
0.10-0.19	1	GA05
0.00-0.09	0	
-0.10- -0.01	1	GA03

Discrimination index range = -0.03 (Item GA03) to 0.82 (Item GA37)
Mean = 0.51
SD = 0.18

Table 3(c) ELTS Test (M1-General Academic): Items Grouped by Unbiased Point Biserial

Point Biserial Interval	No.of Items	Item Names
0.90-1.00	0	
0.80-0.89	0	
0.70-0.79	0	
0.60-0.69	1	GA37
0.50-0.59	6	GA20 GA24 GA34 GA36 GA39 GA40
0.40-0.49	13	GA09 GA14 GA18 GA21 GA23 GA25 GA28 GA29 GA31 GA32 GA33 GA35 GA38
0.30-0.39	10	GA02 GA04 GA07 GA10 GA15 GA16 GA19 GA22 GA27 GA30
0.20-0.29	7	GA06 GA08 GA11 GA12 GA13 GA17 GA26
0.10-0.19	2	GA01 GA05
0.00-0.09	0	
-0.10- -0.01	1	GA03

Point biserial range = -0.08 (Item GA03) to 0.61 (Item GA37)
Mean = 0.38
SD = 0.14

Table 4(a) ELTS Test (M1-Life Sciences): Items Grouped by Facility Value

Facility Value Interval	No.of Items	Item Names
0.90-1.00	2	LS05 LS06
0.80-0.89	9	LS01 LS02 LS07 LS09 LS13 LS15 LS18 LS23 LS24
0.70-0.79	5	LS03 LS11 LS12 LS19 LS21
0.60-0.69	4	LS10 LS20 LS25 LS29
0.50-0.59	8	LS04 LS16 LS17 LS26 LS30 LS31 LS33 LS39
0.40-0.49	7	LS14 LS22 LS27 LS28 LS32 LS37 LS40
0.30-0.39	4	LS08 LS34 LS36 LS38
0.20-0.29	0	
0.10-0.19	0	
0.00-0.09	1	LS35

Facility value range = 0.05 (Item LS35) to 0.95 (Item LS05)
Mean = 0.62
SD = 0.20

Table 4(b) ELTS Test (M1-Life Sciences): Items Grouped by Discrimination Index

Discrimination Index Interval	No.of Items	Item Names
0.90-1.00	0	
0.80-0.89	0	
0.70-0.79	2	LS37 LS39
0.60-0.69	7	LS27 LS29 LS31 LS32 LS33 LS38 LS40
0.50-0.59	4	LS26 LS28 LS30 LS34
0.40-0.49	3	LS17 LS25 LS36
0.30-0.39	7	LS07 LS10 LS14 LS20 LS21 LS22 LS24
0.20-0.29	11	LS01 LS03 LS09 LS11 LS12 LS13 LS15 LS16 LS18 LS19 LS23
0.10-0.19	3	LS02 LS05 LS08
0.00-0.09	3	LS04 LS06 LS35

Discrimination index range = 0.01 (Item LS35) to 0.75 (Item LS39)
Mean = 0.38
SD = 0.20

Table 4(c) ELTS Test (M1-Life Sciences): Items Grouped by Unbiased Point Biserial

Point Biserial Interval	No.of Items	Item Names
0.90-1.00	0	
0.80-0.89	0	
0.70-0.79	0	
0.60-0.69	0	
0.50-0.59	2	LS31 LS39
0.40-0.49	8	LS27 LS29 LS30 LS32 LS33 LS37 LS38 LS40
0.30-0.39	5	LS23 LS24 LS26 LS28 LS34
0.20-0.29	13	LS03 LS05 LS07 LS10 LS12 LS13 LS15 LS17 LS19 LS20 LS21 LS25 LS36
0.10-0.19	9	LS01 LS02 LS06 LS09 LS11 LS14 LS16 LS18 LS22
0.00-0.09	2	LS04 LS08
-0.10- -0.01	1	LS35

Point biserial range = -0.01 (Item LS35) to 0.50 (Items LS31, LS39)
Mean = 0.27
SD = 0.13

Table 5(a) ELTS Test (M1-Medicine): Items Grouped by Facility Value

Facility Value Interval	No.of Items	Item Names
0.90-1.00	6	ME02ME11ME12ME17ME18ME33
0.80-0.89	12	ME01ME03ME07ME08ME09ME10ME14ME16ME19ME20ME22ME31
0.70-0.79	6	ME04ME15ME29ME34ME37ME38
0.60-0.69	4	ME05ME23ME27ME36
0.50-0.59	8	ME24ME25ME26ME28ME30ME32ME39ME40
0.40-0.49	1	ME21
0.30-0.39	3	ME06ME13ME35
0.20-0.29	0	
0.10-0.19	0	
0.00-0.09	0	

Facility value range = 0.30 (Item ME13) to 0.94 (Item ME17)
Mean = 0.71
SD = 0.17

Table 5(b) ELTS Test (M1-Medicine): Items Grouped by Discrimination Index

Discrimination Index Interval	No.of Items	Item Names
0.90-1.00	0	
0.80-0.89	0	
0.70-0.79	1	ME27
0.60-0.69	3	ME23ME36ME39
0.50-0.59	6	ME21ME24ME28ME29ME30ME34
0.40-0.49	8	ME19ME25ME26ME31ME32ME37ME38ME40
0.30-0.39	5	ME03ME04ME06ME10ME15
0.20-0.29	12	ME01ME02ME05ME08ME09ME13ME14ME16ME17ME20ME33ME35
0.10-0.19	5	ME07ME11ME12ME18ME21
0.00-0.09	0	

Discrimination index range = 0.13 (Item ME11) to 0.74 (Item ME27)
Mean = 0.38
SD = 0.16

Table 5(c) : ELTS Test (M1-Medicine): Items Grouped by Unbiased Point Biserial

Point Biserial Interval	No.of Items	Item Names
0.90-1.00	0	
0.80-0.89	0	
0.70-0.79	0	
0.60-0.69	0	
0.50-0.59	2	ME31ME34
0.40-0.49	14	ME02ME03ME09ME10ME17ME19ME23ME27ME29ME30ME33ME36 ME38ME39
0.30-0.39	9	ME01ME15ME16ME21ME24ME28ME32ME37ME40
0.20-0.29	9	ME06ME07ME08ME11ME12ME14ME18ME25ME26
0.10-0.19	5	ME04ME05ME13ME22ME35
0.00-0.09	1	ME20

Point biserial range = 0.08 (Item ME20) to 0.54 (Item ME34)
Mean = 0.34
SD = 0.11

Table 6(a) ELTS Test (M1-Physical Sciences): Items Grouped by Facility Value

Facility Value Interval	No.of Items	Item Names
0.90-1.00	7	PS01 PS08 PS09 PS11 PS16 PS19 PS22
0.80-0.89	11	PS03 PS05 PS17 PS18 PS20 PS23 PS24 PS25 PS27 PS30 PS34
0.70-0.79	10	PS02 PS10 PS14 PS15 PS21 PS26 PS28 PS33 PS36 PS37
0.60-0.69	7	PS06 PS12 PS29 PS31 PS32 PS38 PS39
0.50-0.59	2	PS13 PS35
0.40-0.49	1	PS07
0.30-0.39	1	PS04
0.20-0.29	0	
0.10-0.19	1	PS40
0.00-0.09	0	

Facility value range = 0.15 (Item PS40) to 0.96 (Item PS09)
Mean = 0.75
SD = 0.16

Table 6(b) ELTS Test (M1-Physical Sciences): Items Grouped by Discrimination Index

Discrimination Index Interval	No.of Items	Item Names
0.90-1.00	0	
0.80-0.89	0	
0.70-0.79	2	PS13 PS38
0.60-0.69	5	PS07 PS29 PS31 PS33 PS37
0.50-0.59	8	PS02 PS04 PS14 PS21 PS26 PS35 PS36 PS39
0.40-0.49	7	PS03 PS05 PS10 PS18 PS28 PS32 PS34
0.30-0.39	7	PS06 PS12 PS15 PS23 PS25 PS27 PS30
0.20-0.29	7	PS08 PS16 PS17 PS19 PS20 PS22 PS24
0.10-0.19	3	PS01 PS09 PS11
0.00-0.09	1	PS40

Discrimination index range = 0.03 (Item PS40) to 0.78 (Item PS13)
Mean = 0.42
SD = 0.18

Table 6(c) ELTS Test (M1-Physical Sciences): Items Grouped by Unbiased Point Biserial

Point Biserial Interval	No.of Items	Item Names
0.90-1.00	0	
0.80-0.89	0	
0.70-0.79	0	
0.60-0.69	0	
0.50-0.59	10	PS02 PS09 PS13 PS16 PS21 PS26 PS29 PS34 PS37 PS38
0.40-0.49	13	PS03 PS05 PS07 PS10 PS11 PS19 PS23 PS27 PS28 PS30 PS31 PS33 PS36
0.30-0.39	15	PS01 PS04 PS08 PS12 PS14 PS15 PS17 PS18 PS20 PS22 PS24 PS25 PS32 PS35 PS39
0.20-0.29	1	PS06
0.10-0.19	0	
0.00-0.09	1	PS40

Point biserial range = 0.05 (Item PS40) to 0.54 (Items PS09, PS13)
Mean = 0.41
SD = 0.10

Table 7(a) ELTS Test (M1-Social Studies): Items Grouped by Facility Value

Facility Value Interval	No.of Items	Item Names
0.90-1.00	2	SS16 SS17
0.80-0.89	3	SS09 SS15 SS18
0.70-0.79	1	SS27
0.60-0.69	8	SS02 SS04 SS06 SS07 SS13 SS20 SS28 SS37
0.50-0.59	11	SS03 SS10 SS11 SS14 SS23 SS24 SS25 SS26 SS31 SS33 SS35
0.40-0.49	7	SS01 SS21 SS22 SS30 SS32 SS34 SS39
0.30-0.39	6	SS05 SS08 SS12 SS29 SS36 SS38
0.20-0.29	1	SS40
0.10-0.19	1	SS19
0.00-0.09	0	

Facility value range = 0.16 (Item SS19) to 0.92 (Item SS17)
Mean = 0.55
SD = 0.17

Table 7(b) ELTS Test (M1-Social Studies): Items Grouped by Discrimination Index

Discrimination Index Interval	No.of Items	Item Names
0.90-1.00	0	
0.80-0.89	0	
0.70-0.79	0	
0.60-0.69	6	SS11 SS28 SS30 SS31 SS32 SS33
0.50-0.59	7	SS01 SS12 SS23 SS27 SS29 SS37 SS38
0.40-0.49	11	SS02 SS03 SS05 SS07 SS08 SS09 SS13 SS21 SS34 SS35 SS39
0.30-0.39	5	SS10 SS18 SS20 SS25 SS26
0.20-0.29	8	SS04 SS06 SS14 SS15 SS22 SS24 SS36 SS40
0.10-0.19	2	SS16 SS17
0.00-0.09	0	
-0.10- -0.01	1	SS19

Discrimination index range = -0.01 (Item SS19) to 0.65 (Items SS32, SS33)
Mean = 0.41
SD = 0.16

Table 7(c) ELTS Test (M1-Social Studies): Items Grouped by Unbiased Point Biserial

Point Biserial Interval	No.of Items	Item Names
0.90-1.00	0	
0.80-0.89	0	
0.70-0.79	0	
0.60-0.69	0	
0.50-0.59	0	
0.40-0.49	7	SS11 SS28 SS30 SS31 SS32 SS33 SS37
0.30-0.39	16	SS01 SS02 SS03 SS05 SS07 SS09 SS12 SS13 SS18 SS23 SS27 SS29 SS34 SS35 SS38 SS39
0.20-0.29	10	SS06 SS08 SS10 SS15 SS16 SS17 SS20 SS21 SS25 SS26
0.10-0.19	6	SS04 SS14 SS22 SS24 SS36 SS40
0.00-0.09	0	
-0.10- -0.01	1	SS19

Point biserial range = -0.04 (Item SS19) to 0.48 (Item SS33)
Mean = 0.29
SD = 0.11

Table 8(a) ELTS Test (M1-Technology): Items Grouped by Facility Value

Facility Value Interval	No.of Items	Item Names
0.90-1.00	6	TN03 TN04 TN06 TN07 TN08 TN32
0.80-0.89	9	TN01 TN02 TN09 TN14 TN15 TN16 TN25 TN29 TN34
0.70-0.79	11	TN05 TN10 TN18 TN19 TN20 TN24 TN26 TN30 TN31 TN36 TN37
0.60-0.69	10	TN11 TN12 TN17 TN22 TN23 TN27 TN28 TN33 TN38 TN40
0.50-0.59	2	TN21 TN35
0.40-0.49	2	TN13 TN39
0.30-0.39	0	
0.20-0.29	0	
0.10-0.19	0	
0.00-0.09	0	

Facility value range = 0.42 (Item TN39) to 0.94 (Item TN06)
Mean = 0.74
SD = 0.13

Table 8(b) ELTS Test (M1-Technology): Items Grouped by Discrimination Index

Discrimination Index Interval	No.of Items	Item Names
0.90-1.00	0	
0.80-0.89	0	
0.70-0.79	0	
0.60-0.69	2	TN23 TN33
0.50-0.59	8	TN10 TN17 TN18 TN31 TN35 TN36 TN37 TN40
0.40-0.49	9	TN02 TN05 TN12 TN16 TN19 TN20 TN28 TN34 TN39
0.30-0.39	9	TN04 TN11 TN14 TN15 TN21 TN22 TN25 TN26 TN29
0.20-0.29	10	TN01 TN03 TN07 TN08 TN09 TN24 TN27 TN30 TN32 TN38
0.10-0.19	2	TN06 TN13
0.00-0.09	0	

Discrimination index range = 0.14 (Item TN06) to 0.62 (Item TN23)
Mean = 0.38
SD = 0.13

Table 8(c) ELTS Test (M1-Technology): Items Grouped by Unbiased Point Biserial

Point Biserial Interval	No.of Items	Item Names
0.90-1.00	0	
0.80-0.89	0	
0.70-0.79	0	
0.60-0.69	3	TN03 TN15 TN32
0.50-0.59	8	TN04 TN07 TN08 TN16 TN25 TN31 TN34 TN36
0.40-0.49	11	TN02 TN05 TN10 TN14 TN17 TN18 TN20 TN23 TN26 TN29 TN33
0.30-0.39	10	TN01 TN09 TN11 TN12 TN19 TN21 TN30 TN35 TN37 TN40
0.20-0.29	2	TN06 TN38
0.10-0.19	4	TN22 TN24 TN28 TN39
0.00-0.09	1	TN27
-0.10- -0.01	1	TN13

Point biserial range = -0.10 (Item TN13) to 0.63 (Item TN15)
Mean = 0.39
SD = 0.16

H.4 Facility Values for G1 & G2 from High- & Low-Scoring Subgroups

G1 subtest			G2 subtest		
ITEM NAME	500 Highest scorers	500 Lowest scorers	ITEM NAME	500 Highest scorers	500 Lowest scorers
G101	1.00	0.83	G201	0.98	0.85
G102	0.99	0.80	G202	0.91	0.52
G103	0.93	0.65	G203	0.79	0.41
G104	0.99	0.75	G204	0.83	0.36
G105	0.95	0.70	G205	0.96	0.68
G106	0.94	0.76	G206	0.92	0.50
G107	0.95	0.72	G207	0.89	0.51
G108	0.95	0.74	G208	0.90	0.51
G109	0.78	0.42	G209	0.86	0.42
G110	0.86	0.33	G210	0.89	0.53
G111	0.55	0.28	G211	0.85	0.46
G112	0.72	0.29	G212	0.95	0.59
G113	0.99	0.70	G213	0.63	0.21
G114	0.96	0.54	G214	0.81	0.31
G115	0.92	0.54	G215	0.95	0.64
G116	0.93	0.59	G216	0.79	0.30
G117	0.83	0.32	G217	0.96	0.73
G118	0.89	0.45	G218	0.99	0.76
G119	0.97	0.79	G219	0.97	0.74
G120	0.97	0.76	G220	0.99	0.86
G121	0.96	0.55	G221	0.95	0.65
G122	0.98	0.66	G222	0.92	0.45
G123	0.96	0.55	G223	0.95	0.54
G124	0.86	0.46	G224	0.96	0.65
G125	0.62	0.27	G225	0.74	0.32
G126	0.73	0.21	G226	0.92	0.72
G127	0.98	0.70	G227	0.08	0.13
G128	0.87	0.52	G228	0.80	0.35
G129	0.94	0.55	G229	0.82	0.40
G130	0.90	0.48	G230	0.52	0.15
G131	0.86	0.38	G231	0.95	0.69
G132	0.79	0.34	G232	0.86	0.48
G133	0.99	0.54	G233	0.76	0.40
G134	0.96	0.44	G234	0.81	0.41
G135	0.84	0.22	G235	0.40	0.19
G136	0.96	0.43			
G137	0.92	0.21			
G138	0.78	0.13			
G139	0.90	0.38			
G140	0.93	0.28			

APPENDIX I
RASCH STATISTICS FOR ELTS TEST

I.1 ELTS Subtests: Raw Scores, Rasch Ability Estimates & Standard Errors

(I) G1 (Reading)

TABLE 1
(All measurable persons included)

RAW SCORE	FREQ. COUNT	ABILITY ESTIM.	STANDARD ERROR
39	40	3.94	1.03
38	58	3.22	0.74
37	61	2.78	0.62
36	84	2.45	0.55
35	75	2.19	0.50
34	74	1.96	0.47
33	56	1.76	0.44
32	71	1.58	0.42
31	58	1.41	0.41
30	86	1.26	0.39
29	62	1.11	0.38
28	88	0.98	0.37
27	76	0.84	0.37
26	74	0.71	0.36
25	72	0.59	0.36
24	72	0.47	0.35
23	51	0.35	0.35
22	61	0.23	0.35
21	42	0.11	0.35
20	31	-0.00	0.35
19	39	-0.12	0.35
18	26	-0.24	0.35
17	31	-0.36	0.35
16	32	-0.47	0.35
15	16	-0.60	0.35
14	13	-0.72	0.36
13	8	-0.85	0.37
12	8	-0.98	0.37
11	3	-1.12	0.38
10	4	-1.26	0.39
9	2	-1.41	0.40
8	3	-1.58	0.42
7	3	-1.76	0.44
6	0	-1.96	0.47
5	0	-2.18	0.50
4	0	-2.44	0.55
3	0	-2.77	0.62
2	0	-3.21	0.74
1	0	-3.93	1.03

No. of persons = 1,480

Mean ability = 1.16
Sd ability = 1.01
Group ability range:-1.76 to 3.94

Person separability index = 0.83
No. of person strata = 3.28

TABLE 2
(32 misfitting persons excluded)

RAW SCORE	FREQ. COUNT	ABILITY ESTIM.	STANDARD ERROR
39	40	3.97	1.03
38	58	3.25	0.75
37	61	2.80	0.62
36	84	2.48	0.55
35	75	2.21	0.50
34	74	1.98	0.47
33	56	1.78	0.44
32	71	1.60	0.42
31	58	1.43	0.41
30	86	1.28	0.39
29	62	1.13	0.38
28	88	0.99	0.38
27	75	0.85	0.37
26	72	0.72	0.36
25	71	0.60	0.36
24	70	0.47	0.35
23	49	0.35	0.35
22	58	0.23	0.35
21	36	0.12	0.35
20	30	-0.00	0.35
19	36	-0.12	0.35
18	25	-0.24	0.35
17	30	-0.36	0.35
16	32	-0.48	0.35
15	11	-0.60	0.36
14	11	-0.73	0.36
13	6	-0.86	0.37
12	8	-0.99	0.37
11	3	-1.13	0.38
10	4	-1.28	0.39
9	2	-1.43	0.41
8	3	-1.60	0.42
7	3	-1.78	0.44
6	0	-1.98	0.47
5	0	-2.20	0.50
4	0	-2.47	0.55
3	0	-2.80	0.62
2	0	-3.24	0.75
1	0	-3.96	1.03

No. of persons = 1,448

Mean ability = 1.20
SD ability =1.01
Group ability range:-1.78 to 3.97

Person separability index = 0.83
No. of person strata = 3.25

(ii) G2 (Listening)

TABLE 1
(All measurable persons included)

RAW SCORE	FREQ. COUNT	ABILITY ESTIM.	STANDARD ERROR
34	11	4.08	1.09
33	37	3.28	0.80
32	41	2.77	0.67
31	41	2.40	0.59
30	83	2.10	0.54
29	77	1.84	0.50
28	78	1.61	0.47
27	107	1.41	0.45
26	118	1.23	0.43
25	105	1.05	0.41
24	104	0.89	0.40
23	99	0.74	0.39
22	96	0.59	0.39
21	95	0.45	0.38
20	63	0.31	0.38
19	73	0.17	0.38
18	63	0.03	0.37
17	51	-0.10	0.37
16	45	-0.24	0.37
15	29	-0.37	0.38
14	20	-0.51	0.38
13	18	-0.65	0.38
12	20	-0.80	0.39
11	12	-0.95	0.40
10	6	-1.10	0.41
9	5	-1.27	0.42
8	3	-1.44	0.43
7	1	-1.63	0.45
6	0	-1.84	0.48
5	1	-2.08	0.51
4	1	-2.36	0.56
3	0	-2.69	0.63
2	0	-3.15	0.76
1	0	-3.88	1.04

No. of persons = 1,503

Mean ability = 0.95
Sd ability = 0.85
Group ability range:-2.36 to 4.08

Person separability index = 0.78
No. of person strata = 2.83

TABLE 2
(22 misfitting persons excluded)

RAW SCORE	FREQ. COUNT	ABILITY ESTIM.	STANDARD ERROR
34	11	4.11	1.09
33	37	3.30	0.80
32	41	2.79	0.67
31	41	2.42	0.59
30	83	2.11	0.54
29	77	1.85	0.50
28	78	1.63	0.47
27	107	1.42	0.45
26	118	1.24	0.43
25	104	1.06	0.42
24	103	0.90	0.40
23	99	0.74	0.40
22	96	0.59	0.39
21	95	0.45	0.38
20	63	0.31	0.38
19	69	0.17	0.38
18	62	0.03	0.37
17	49	-0.10	0.37
16	40	-0.24	0.38
15	26	-0.38	0.38
14	18	-0.52	0.38
13	18	-0.66	0.39
12	19	-0.80	0.39
11	11	-0.95	0.40
10	5	-1.11	0.41
9	5	-1.28	0.42
8	3	-1.46	0.44
7	1	-1.65	0.46
6	0	-1.86	0.48
5	1	-2.10	0.52
4	1	-2.38	0.56
3	0	-2.72	0.63
2	0	-3.17	0.76
1	0	-3.91	1.04

No. of persons = 1,481

Mean ability = 0.97
SD ability = 0.85
Group ability range:-2.38 to 4.11

Person separability index = 0.78
No. of person strata = 2.81

(iii) M1 (General Academic)

TABLE 1
(All measurable persons included)

RAW SCORE	FREQ. COUNT	ABILITY ESTIM.	STANDARD ERROR
39	2	3.80	1.02
38	3	3.09	0.74
37	6	2.65	0.61
36	12	2.34	0.54
35	6	2.08	0.49
34	8	1.86	0.46
33	8	1.67	0.43
32	12	1.50	0.41
31	7	1.34	0.39
30	10	1.20	0.38
29	9	1.06	0.37
28	16	0.93	0.36
27	12	0.80	0.36
26	13	0.68	0.35
25	12	0.57	0.34
24	8	0.45	0.34
23	16	0.34	0.34
22	17	0.23	0.34
21	17	0.12	0.34
20	23	0.01	0.34
19	16	-0.10	0.34
18	19	-0.21	0.34
17	16	-0.32	0.34
16	21	-0.44	0.34
15	18	-0.55	0.35
14	14	-0.67	0.35
13	18	-0.79	0.36
12	19	-0.92	0.36
11	14	-1.05	0.37
10	10	-1.19	0.38
9	5	-1.34	0.40
8	3	-1.50	0.41
7	5	-1.68	0.43
6	3	-1.87	0.46
5	4	-2.09	0.50
4	0	-2.35	0.54
3	0	-2.67	0.62
2	0	-3.11	0.74
1	0	-3.83	1.03

No. of persons = 402

Mean ability = 0.19
Sd ability = 1.00
Group ability range:-2.09 to 3.80

Person separability index = 0.87
No. of person strata = 3.78

TABLE 2
(10 misfitting persons excluded)

RAW SCORE	FREQ. COUNT	ABILITY ESTIM.	STANDARD ERROR
39	2	3.82	1.02
38	3	3.10	0.74
37	6	2.67	0.61
36	12	2.35	0.54
35	6	2.09	0.49
34	8	1.88	0.46
33	8	1.68	0.43
32	12	1.51	0.41
31	7	1.35	0.40
30	10	1.21	0.38
29	9	1.07	0.37
28	16	0.94	0.36
27	12	0.81	0.36
26	13	0.69	0.35
25	11	0.57	0.35
24	8	0.46	0.34
23	15	0.34	0.34
22	15	0.23	0.34
21	16	0.12	0.34
20	22	0.01	0.34
19	16	-0.10	0.34
18	18	-0.21	0.34
17	14	-0.33	0.34
16	20	-0.44	0.34
15	18	-0.56	0.35
14	14	-0.68	0.35
13	18	-0.80	0.36
12	19	-0.93	0.37
11	14	-1.06	0.38
10	10	-1.20	0.39
9	5	-1.35	0.40
8	3	-1.51	0.42
7	5	-1.69	0.44
6	3	-1.88	0.46
5	4	-2.11	0.50
4	0	-2.37	0.54
3	0	-2.69	0.62
2	0	-3.13	0.74
1	0	-3.85	1.03

No. of persons = 392

Mean ability = 0.19
SD ability = 1.02
Group ability range:-2.11 to 3.82

Person separability index = 0.87
No. of person strata = 3.79

(iv) M1 (Life Sciences)

TABLE 1
(All measurable persons included)

RAW SCORE	FREQ. COUNT	ABILITY ESTIM.	STANDARD ERROR
39	0	4.32	1.11
38	0	3.48	0.81
37	3	2.97	0.67
36	10	2.60	0.58
35	7	2.30	0.53
34	7	2.05	0.49
33	14	1.84	0.46
32	14	1.65	0.43
31	13	1.47	0.42
30	17	1.31	0.40
29	15	1.16	0.39
28	30	1.01	0.38
27	26	0.87	0.37
26	11	0.74	0.37
25	23	0.61	0.36
24	17	0.48	0.36
23	26	0.36	0.36
22	15	0.24	0.35
21	18	0.11	0.35
20	26	-0.01	0.35
19	25	-0.13	0.35
18	16	-0.25	0.36
17	5	-0.38	0.36
16	12	-0.50	0.36
15	8	-0.63	0.37
14	6	-0.76	0.37
13	3	-0.90	0.38
12	0	-1.04	0.38
11	2	-1.19	0.39
10	2	-1.34	0.40
9	1	-1.51	0.42
8	1	-1.68	0.43
7	0	-1.87	0.46
6	0	-2.08	0.48
5	0	-2.32	0.52
4	0	-2.60	0.57
3	0	-2.95	0.64
2	0	-3.41	0.76
1	0	-4.15	1.05

No. of persons = 373

Mean ability = 0.63
Sd ability = 0.76
Group ability range:-1.68 to 2.97

Person separability index = 0.79
No. of person strata = 2.92

TABLE 2
(18 misfitting persons excluded)

RAW SCORE	FREQ. COUNT	ABILITY ESTIM.	STANDARD ERROR
39	0	4.41	1.13
38	0	3.56	0.82
37	3	3.03	0.67
36	10	2.65	0.59
35	7	2.35	0.53
34	7	2.10	0.49
33	14	1.88	0.46
32	14	1.69	0.44
31	13	1.51	0.42
30	17	1.35	0.41
29	15	1.19	0.39
28	29	1.04	0.38
27	24	0.90	0.38
26	11	0.77	0.37
25	21	0.63	0.37
24	16	0.50	0.36
23	24	0.38	0.36
22	14	0.25	0.36
21	16	0.12	0.36
20	24	-0.00	0.36
19	23	-0.13	0.36
18	15	-0.25	0.36
17	5	-0.38	0.36
16	11	-0.51	0.37
15	8	-0.64	0.37
14	6	-0.78	0.38
13	2	-0.92	0.38
12	0	-1.06	0.39
11	2	-1.22	0.40
10	2	-1.38	0.41
9	1	-1.54	0.42
8	1	-1.73	0.44
7	0	-1.92	0.46
6	0	-2.14	0.49
5	0	-2.39	0.52
4	0	-2.68	0.57
3	0	-3.03	0.65
2	0	-3.50	0.77
1	0	-4.26	1.06

No. of persons = 355

Mean ability = 0.67
SD ability = 0.79
Group ability range:-1.73 to 3.03

Person separability index =0.80
No. of person strata = 2.98

(v) M1 (Medicine)

TABLE 1
(All measurable persons included)

RAW SCORE	FREQ. COUNT	ABILITY ESTIM.	STANDARD ERROR
39	0	4.08	1.04
38	3	3.35	0.75
37	6	2.90	0.63
36	6	2.56	0.56
35	6	2.29	0.51
34	15	2.05	0.48
33	12	1.85	0.45
32	8	1.66	0.43
31	9	1.49	0.41
30	9	1.32	0.40
29	9	1.17	0.39
28	6	1.03	0.38
27	8	0.89	0.38
26	7	0.75	0.37
25	5	0.62	0.37
24	4	0.49	0.36
23	5	0.36	0.36
22	6	0.24	0.36
21	4	0.11	0.36
20	4	-0.01	0.36
19	2	-0.13	0.36
18	0	-0.26	0.36
17	1	-0.38	0.36
16	2	-0.51	0.36
15	0	-0.63	0.36
14	1	-0.76	0.37
13	0	-0.90	0.37
12	0	-1.04	0.38
11	1	-1.18	0.39
10	0	-1.33	0.40
9	3	-1.49	0.41
8	1	-1.66	0.43
7	0	-1.84	0.45
6	0	-2.05	0.47
5	0	-2.28	0.50
4	0	-2.54	0.55
3	0	-2.88	0.62
2	0	-3.32	0.75
1	0	-4.05	1.03

No. of persons = 143

Mean ability = 1.23
Sd ability = 0.91
Group ability range:-1.66 to 3.35

Person separability index = 0.81
No. of person strata = 3.10

TABLE 2
(1 misfitting person excluded)

RAW SCORE	FREQ. COUNT	ABILITY ESTIM.	STANDARD ERROR
39	0	4.09	1.04
38	3	3.36	0.75
37	6	2.91	0.63
36	6	2.57	0.56
35	6	2.30	0.51
34	15	2.07	0.48
33	12	1.86	0.45
32	8	1.67	0.43
31	9	1.50	0.42
30	9	1.33	0.40
29	9	1.18	0.39
28	6	1.03	0.38
27	8	0.89	0.38
26	7	0.76	0.37
25	5	0.62	0.37
24	4	0.49	0.36
23	5	0.37	0.36
22	6	0.24	0.36
21	4	0.12	0.36
20	4	-0.01	0.36
19	2	-0.13	0.36
18	0	-0.26	0.36
17	1	-0.38	0.36
16	2	-0.51	0.36
15	0	-0.64	0.37
14	1	-0.77	0.37
13	0	-0.90	0.37
12	0	-1.04	0.38
11	1	-1.19	0.39
10	0	-1.34	0.40
9	2	-1.50	0.41
8	1	-1.67	0.43
7	0	-1.85	0.45
6	0	-2.06	0.47
5	0	-2.29	0.51
4	0	-2.56	0.55
3	0	-2.89	0.63
2	0	-3.34	0.75
1	0	-4.07	1.03

No. of persons = 142

Mean ability = 1.26
SD ability = 0.88
Group ability range:-1.67 to 3.36

Person separability index = 0.80
No. of person strata = 3.01

(vi) M1 (Physical Sciences)

TABLE 1
(All measurable persons included)

RAW SCORE	FREQ. COUNT	ABILITY ESTIM.	STANDARD ERROR
39	3	4.31	1.09
38	8	3.49	0.80
37	12	2.99	0.67
36	9	2.61	0.59
35	9	2.31	0.53
34	11	2.06	0.49
33	8	1.84	0.46
32	11	1.64	0.44
31	8	1.46	0.42
30	4	1.30	0.41
29	4	1.14	0.39
28	2	0.99	0.38
27	10	0.85	0.38
26	2	0.72	0.37
25	7	0.58	0.36
24	0	0.46	0.36
23	5	0.33	0.36
22	1	0.21	0.35
21	6	0.09	0.35
20	1	-0.04	0.35
19	4	-0.16	0.35
18	0	-0.28	0.35
17	3	-0.40	0.36
16	1	-0.52	0.36
15	1	-0.65	0.36
14	0	-0.78	0.37
13	0	-0.91	0.37
12	0	-1.05	0.38
11	0	-1.19	0.39
10	1	-1.34	0.40
9	0	-1.50	0.41
8	2	-1.67	0.43
7	0	-1.85	0.45
6	0	-2.06	0.47
5	0	-2.29	0.51
4	0	-2.56	0.55
3	0	-2.89	0.63
2	0	-3.34	0.75
1	0	-4.07	1.04

No. of persons = 133

Mean ability = 1.57
Sd ability = 1.10
Group ability range:-1.67 to 4.31

Person separability index = 0.82
No. of person strata = 3.19

TABLE 2
(1 misfitting person excluded)

RAW SCORE	FREQ. COUNT	ABILITY ESTIM.	STANDARD ERROR
39	3	4.33	1.09
38	8	3.51	0.80
37	12	3.00	0.67
36	9	2.62	0.59
35	9	2.32	0.53
34	11	2.07	0.49
33	8	1.85	0.46
32	11	1.65	0.44
31	8	1.47	0.42
30	4	1.30	0.41
29	4	1.14	0.40
28	2	1.00	0.39
27	10	0.85	0.38
26	2	0.72	0.37
25	6	0.59	0.37
24	0	0.46	0.36
23	5	0.33	0.36
22	1	0.21	0.36
21	6	0.09	0.35
20	1	-0.04	0.35
19	4	-0.16	0.35
18	0	-0.28	0.35
17	3	-0.40	0.36
16	1	-0.53	0.36
15	1	-0.65	0.36
14	0	-0.78	0.37
13	0	-0.92	0.37
12	0	-1.05	0.38
11	0	-1.20	0.39
10	1	-1.35	0.40
9	0	-1.51	0.41
8	2	-1.68	0.43
7	0	-1.86	0.45
6	0	-2.07	0.47
5	0	-2.30	0.51
4	0	-2.57	0.56
3	0	-2.90	0.63
2	0	-3.35	0.75
1	0	-4.08	1.04

No. of persons = 132

Mean ability = 1.58
SD ability = 1.11
Group ability range:-1.68 to 4.33

Person separability index = 0.82
No. of person strata = 3.20

(vii) **M1 (Social Studies)**

TABLE 1
(All measurable persons included)

RAW SCORE	FREQ. COUNT	ABILITY ESTIM.	STANDARD ERROR
39	1	3.90	1.03
38	1	3.18	0.74
37	1	2.74	0.62
36	1	2.41	0.55
35	3	2.15	0.50
34	7	1.93	0.46
33	4	1.73	0.44
32	2	1.56	0.42
31	9	1.40	0.40
30	8	1.25	0.39
29	9	1.11	0.38
28	6	0.97	0.37
27	9	0.84	0.36
26	15	0.72	0.35
25	15	0.60	0.35
24	18	0.48	0.35
23	13	0.37	0.34
22	17	0.25	0.34
21	15	0.14	0.34
20	16	0.03	0.34
19	16	-0.09	0.34
18	10	-0.20	0.34
17	9	-0.32	0.35
16	7	-0.43	0.35
15	20	-0.55	0.35
14	5	-0.68	0.36
13	4	-0.81	0.37
12	7	-0.94	0.37
11	5	-1.08	0.38
10	5	-1.23	0.40
9	1	-1.38	0.41
8	2	-1.56	0.43
7	0	-1.74	0.45
6	3	-1.95	0.48
5	0	-2.19	0.51
4	0	-2.46	0.56
3	0	-2.81	0.63
2	0	-3.26	0.76
1	0	-4.01	1.04

No. of persons = 264

Mean ability = 0.28
Sd ability = 0.79
Group ability range:-1.95 to 3.90

Person separability index = 0.82
No. of person strata = 3.15

TABLE 2
(7 misfitting persons excluded)

RAW SCORE	FREQ. COUNT	ABILITY ESTIM.	STANDARD ERROR
39	1	3.93	1.03
38	1	3.21	0.75
37	1	2.77	0.62
36	1	2.44	0.55
35	3	2.17	0.50
34	7	1.95	0.46
33	4	1.75	0.44
32	2	1.57	0.42
31	9	1.41	0.40
30	8	1.26	0.39
29	9	1.12	0.38
28	6	0.98	0.37
27	9	0.85	0.36
26	15	0.73	0.36
25	15	0.61	0.35
24	17	0.49	0.35
23	13	0.37	0.34
22	16	0.26	0.34
21	15	0.14	0.34
20	16	0.03	0.34
19	15	-0.08	0.34
18	10	-0.20	0.34
17	8	-0.32	0.35
16	7	-0.44	0.35
15	19	-0.56	0.36
14	5	-0.68	0.36
13	4	-0.81	0.37
12	6	-0.95	0.38
11	5	-1.09	0.39
10	4	-1.24	0.40
9	1	-1.40	0.41
8	2	-1.57	0.43
7	0	-1.76	0.45
6	3	-1.97	0.48
5	0	-2.21	0.52
4	0	-2.49	0.57
3	0	-2.84	0.64
2	0	-3.31	0.76
1	0	-4.06	1.05

No. of persons = 257

Mean ability = 0.30
SD ability = 0.80
Group ability range:-1.97 to 3.93

Person separability index = 0.82
No. of person strata = 3.17

(viii) **M1 (Technology)**

TABLE 1
(All measurable persons included)

RAW SCORE	FREQ. COUNT	ABILITY ESTIM.	STANDARD ERROR
39	4	3.92	1.03
38	4	3.20	0.74
37	8	2.76	0.62
36	12	2.44	0.55
35	15	2.17	0.50
34	13	1.95	0.46
33	20	1.75	0.44
32	15	1.58	0.42
31	10	1.41	0.40
30	13	1.26	0.39
29	14	1.12	0.38
28	11	0.98	0.37
27	3	0.85	0.36
26	7	0.72	0.36
25	7	0.60	0.35
24	4	0.48	0.35
23	0	0.36	0.35
22	5	0.25	0.35
21	3	0.13	0.35
20	0	0.01	0.35
19	3	-0.10	0.35
18	2	-0.22	0.35
17	1	-0.34	0.35
16	0	-0.46	0.35
15	1	-0.58	0.36
14	1	-0.71	0.36
13	0	-0.84	0.37
12	0	-0.97	0.37
11	0	-1.11	0.38
10	1	-1.26	0.39
9	0	-1.41	0.41
8	1	-1.58	0.42
7	2	-1.76	0.44
6	4	-1.96	0.47
5	0	-2.19	0.50
4	0	-2.46	0.55
3	0	-2.79	0.62
2	0	-3.23	0.75
1	0	-3.96	1.03

No. of persons = 184

Mean ability = 1.37
Sd ability = 0.97
Group ability range:-1.96 to 3.92

Person separability index = 0.82
No. of person strata = 3.13

TABLE 2
(2 misfitting persons excluded)

RAW SCORE	FREQ. COUNT	ABILITY ESTIM.	STANDARD ERROR
39	4	3.93	1.03
38	4	3.21	0.74
37	8	2.78	0.62
36	12	2.45	0.55
35	15	2.19	0.50
34	13	1.96	0.46
33	20	1.77	0.44
32	15	1.59	0.42
31	10	1.43	0.40
30	13	1.27	0.39
29	14	1.13	0.38
28	11	0.99	0.37
27	3	0.86	0.36
26	7	0.73	0.36
25	7	0.61	0.35
24	4	0.49	0.35
23	0	0.37	0.35
22	4	0.25	0.35
21	3	0.13	0.35
20	0	0.02	0.35
19	3	-0.10	0.35
18	2	-0.22	0.35
17	1	-0.34	0.35
16	0	-0.46	0.35
15	0	-0.58	0.36
14	1	-0.71	0.36
13	0	-0.84	0.37
12	0	-0.98	0.38
11	0	-1.12	0.39
10	1	-1.27	0.40
9	0	-1.42	0.41
8	1	-1.59	0.43
7	2	-1.78	0.45
6	4	-1.98	0.47
5	0	-2.21	0.51
4	0	-2.48	0.55
3	0	-2.81	0.63
2	0	-3.26	0.75
1	0	-3.99	1.03

No. of persons = 182

Mean ability = 1.40
SD ability = 0.97
Group ability range:-1.98 to 3.93

Person separability index = 0.81
No. of person strata = 3.12

I.2 ELTS Subtests: Final Item Difficulty Estimates & Standard Errors

(i) G1 (Reading)

(32 misfitting persons excluded)

ITEM NAME	ITEM DIFFICULTY	STANDARD ERROR
G101	-1.90	0.11
G102	-1.69	0.10
G103	-0.51	0.07
G104	-1.24	0.09
G105	-0.70	0.08
G106	-1.02	0.08
G107	-0.66	0.07
G108	-0.99	0.08
G109	0.79	0.06
G110	0.65	0.06
G111	1.69	0.06
G112	1.24	0.06
G113	-1.11	0.08
G114	-0.20	0.07
G115	-0.14	0.07
G116	-0.22	0.07
G117	0.77	0.06
G118	0.25	0.06
G119	-1.44	0.09
G120	-1.20	0.09
G121	-0.33	0.07
G122	-0.70	0.08
G123	-0.26	0.07
G124	0.35	0.06
G125	1.60	0.06
G126	1.52	0.06
G127	-0.98	0.08
G128	0.21	0.06
G129	-0.20	0.07
G130	0.16	0.06
G131	0.59	0.06
G132	0.82	0.06
G133	-0.49	0.07
G134	0.02	0.06
G135	1.25	0.06
G136	0.12	0.06
G137	1.02	0.06
G138	1.73	0.06
G139	0.49	0.06
G140	0.71	0.06

Items calibrated on 1,448 persons
Mean item difficulty = 0.00
SD item difficulty = 0.96
Difficulty range: -1.90 to 1.73

(ii) G2 (Listening)

(22 misfitting persons excluded)

ITEM NAME	ITEM DIFFICULTY	STANDARD ERROR
G201	-1.93	0.11
G202	-0.12	0.06
G203	0.55	0.06
G204	0.65	0.06
G205	-0.90	0.07
G206	-0.19	0.06
G207	-0.06	0.06
G208	-0.14	0.06
G209	0.29	0.06
G210	-0.22	0.06
G211	0.15	0.06
G212	-0.64	0.07
G213	1.37	0.06
G214	0.71	0.06
G215	-0.66	0.07
G216	0.76	0.06
G217	-1.16	0.08
G218	-1.63	0.09
G219	-1.17	0.08
G220	-2.27	0.12
G221	-0.75	0.07
G222	0.03	0.06
G223	-0.33	0.06
G224	-0.87	0.07
G225	0.76	0.06
G226	-0.85	0.07
G227	3.54	0.09
G228	0.61	0.06
G229	0.46	0.06
G230	1.94	0.06
G231	-0.97	0.08
G232	0.12	0.06
G233	0.55	0.06
G234	0.45	0.06
G235	1.96	0.06

Items calibrated on 1,481 persons
Mean item difficulty = 0.00
SD item difficulty = 1.14
Difficulty range: -2.27 to 3.54

(iii) M1 (General Academic)

(10 misfitting persons excluded)

ITEM NAME	ITEM DIFFICULTY	STANDARD ERROR
GA01	-1.24	0.13
GA02	0.24	0.11
GA03	0.81	0.12
GA04	-0.45	0.12
GA05	0.08	0.11
GA06	-0.65	0.12
GA07	-1.24	0.13
GA08	-1.04	0.12
GA09	0.52	0.11
GA10	-0.77	0.12
GA11	-1.79	0.15
GA12	-1.26	0.13
GA13	0.38	0.11
GA14	-0.09	0.11
GA15	-0.32	0.11
GA16	0.04	0.11
GA17	-0.54	0.12
GA18	0.78	0.12
GA19	0.15	0.11
GA20	-0.38	0.11
GA21	0.73	0.12
GA22	-1.18	0.13
GA23	-0.43	0.11
GA24	-0.21	0.11
GA25	0.86	0.12
GA26	1.27	0.13
GA27	-0.17	0.11
GA28	-0.66	0.12
GA29	0.25	0.11
GA30	0.79	0.12
GA31	-0.10	0.11
GA32	-0.17	0.11
GA33	0.77	0.12
GA34	0.30	0.11
GA35	1.25	0.12
GA36	0.23	0.11
GA37	0.67	0.12
GA38	0.59	0.12
GA39	1.31	0.13
GA40	0.71	0.12

Items calibrated on 392 persons
Mean item difficulty = 0.00
SD item difficulty = 0.77
Difficulty range: -1.79 to 1.31

(iv) M1 (Life Sciences)

(18 misfitting persons excluded)

ITEM NAME	ITEM DIFFICULTY	STANDARD ERROR
LS01	-1.12	0.15
LS02	-1.87	0.19
LS03	-0.38	0.13
LS04	0.65	0.12
LS05	-2.94	0.30
LS06	-2.51	0.25
LS07	-1.07	0.15
LS08	1.51	0.12
LS09	-1.30	0.16
LS10	-0.09	0.12
LS11	-0.43	0.13
LS12	-0.97	0.14
LS13	-1.14	0.15
LS14	0.86	0.12
LS15	-0.91	0.14
LS16	0.35	0.12
LS17	0.26	0.12
LS18	-0.99	0.14
LS19	-0.49	0.13
LS20	-0.22	0.12
LS21	-0.57	0.13
LS22	1.04	0.12
LS23	-1.47	0.17
LS24	-1.37	0.16
LS25	-0.27	0.12
LS26	0.53	0.12
LS27	0.80	0.12
LS28	0.84	0.12
LS29	-0.06	0.12
LS30	0.39	0.12
LS31	0.63	0.12
LS32	0.97	0.12
LS33	0.54	0.12
LS34	1.40	0.12
LS35	4.18	0.28
LS36	1.26	0.12
LS37	0.99	0.12
LS38	1.29	0.12
LS39	0.74	0.12
LS40	0.95	0.12

Items calibrated on 355 persons
Mean item difficulty = 0.00
SD item difficulty = 1.28
Difficulty range: -2.94 to 4.18

(v) M1 (Medicine)

(1 misfitting person excluded)

ITEM NAME	ITEM DIFFICULTY	STANDARD ERROR
ME01	-1.19	0.29
ME02	-1.57	0.33
ME03	-0.47	0.24
ME04	-0.17	0.22
ME05	0.34	0.20
ME06	1.93	0.19
ME07	-0.82	0.26
ME08	-0.96	0.27
ME09	-0.89	0.26
ME10	-0.47	0.24
ME11	-1.57	0.33
ME12	-1.68	0.34
ME13	2.30	0.20
ME14	-0.47	0.24
ME15	-0.12	0.22
ME16	-0.82	0.26
ME17	-2.09	0.40
ME18	-1.68	0.34
ME19	-0.64	0.25
ME20	-0.52	0.24
ME21	1.35	0.18
ME22	-0.70	0.25
ME23	0.49	0.20
ME24	1.02	0.19
ME25	0.95	0.19
ME26	1.12	0.19
ME27	0.71	0.19
ME28	1.15	0.19
ME29	-0.07	0.21
ME30	1.05	0.19
ME31	-0.52	0.24
ME32	0.92	0.19
ME33	-1.28	0.30
ME34	0.26	0.20
ME35	1.83	0.19
ME36	0.75	0.19
ME37	0.10	0.21
ME38	0.10	0.21
ME39	1.22	0.18
ME40	1.05	0.19

Items calibrated on 142 persons
Mean item difficulty = 0.00
SD item difficulty = 1.08
Difficulty range: -2.09 to 2.30

(vi) M1 (Physical Sciences)

(1 misfitting person excluded)

ITEM NAME	ITEM DIFFICULTY	STANDARD ERROR
PS01	-1.72	0.39
PS02	0.35	0.22
PS03	-0.78	0.28
PS04	2.28	0.21
PS05	-0.29	0.25
PS06	1.05	0.20
PS07	1.65	0.20
PS08	-1.23	0.33
PS09	-2.26	0.47
PS10	-0.11	0.24
PS11	-1.72	0.39
PS12	0.59	0.21
PS13	1.21	0.20
PS14	0.31	0.22
PS15	0.31	0.22
PS16	-1.58	0.37
PS17	-0.55	0.27
PS18	-0.29	0.25
PS19	-1.34	0.34
PS20	-0.94	0.30
PS21	0.05	0.23
PS22	-1.23	0.33
PS23	-0.63	0.27
PS24	-0.35	0.25
PS25	-0.29	0.25
PS26	-0.00	0.24
PS27	-0.94	0.30
PS28	-0.00	0.24
PS29	0.85	0.21
PS30	-0.86	0.29
PS31	0.93	0.21
PS32	0.67	0.21
PS33	0.31	0.22
PS34	-0.78	0.28
PS35	1.29	0.20
PS36	0.21	0.23
PS37	0.45	0.22
PS38	0.54	0.22
PS39	1.05	0.20
PS40	3.81	0.27

Items calibrated on 132 persons
Mean item difficulty = 0.00
SD item difficulty = 1.15
Difficulty range: -2.26 to 3.81

(vii) M1 (Social Studies)

(7 misfitting persons excluded)

ITEM NAME	ITEM DIFFICULTY	STANDARD ERROR
SS01	0.37	0.13
SS02	-0.26	0.14
SS03	-0.11	0.14
SS04	-0.59	0.14
SS05	1.19	0.15
SS06	-0.22	0.14
SS07	-0.51	0.14
SS08	0.84	0.14
SS09	-1.30	0.17
SS10	-0.15	0.14
SS11	0.03	0.14
SS12	0.78	0.14
SS13	-0.59	0.14
SS14	-0.02	0.14
SS15	-1.30	0.17
SS16	-2.36	0.23
SS17	-2.59	0.26
SS18	-1.90	0.20
SS19	2.23	0.18
SS20	-0.47	0.14
SS21	0.46	0.14
SS22	0.71	0.14
SS23	0.17	0.13
SS24	0.28	0.13
SS25	-0.08	0.14
SS26	0.05	0.14
SS27	-0.73	0.15
SS28	-0.32	0.14
SS29	0.82	0.14
SS30	0.62	0.14
SS31	0.10	0.13
SS32	0.40	0.13
SS33	-0.13	0.14
SS34	0.74	0.14
SS35	0.07	0.14
SS36	1.25	0.15
SS37	-0.26	0.14
SS38	0.88	0.14
SS39	0.49	0.14
SS40	1.43	0.15

Items calibrated on 257 persons
Mean item difficulty = 0.00
SD item difficulty = 0.95
Difficulty range: -2.59 to 2.23

(viii) M1 (Technology)

(2 misfitting persons excluded)

ITEM NAME	ITEM DIFFICULTY	STANDARD ERROR
TN01	-0.75	0.24
TN02	-0.22	0.21
TN03	-1.63	0.32
TN04	-1.36	0.29
TN05	-0.14	0.20
TN06	-1.96	0.35
TN07	-1.44	0.30
TN08	-1.73	0.33
TN09	-0.75	0.24
TN10	0.40	0.18
TN11	0.79	0.17
TN12	0.90	0.17
TN13	1.56	0.16
TN14	-0.54	0.23
TN15	-1.20	0.28
TN16	-0.49	0.22
TN17	0.55	0.18
TN18	0.30	0.18
TN19	0.16	0.19
TN20	0.02	0.20
TN21	1.22	0.16
TN22	0.90	0.17
TN23	0.67	0.17
TN24	0.49	0.18
TN25	-0.81	0.24
TN26	-0.14	0.20
TN27	0.52	0.18
TN28	0.61	0.17
TN29	-0.18	0.21
TN30	0.16	0.19
TN31	-0.14	0.20
TN32	-1.44	0.30
TN33	0.92	0.17
TN34	-0.64	0.23
TN35	1.24	0.16
TN36	0.23	0.19
TN37	0.13	0.19
TN38	0.98	0.17
TN39	1.87	0.16
TN40	0.95	0.17

Items calibrated on 182 persons
Mean item difficulty = 0.00
SD item difficulty = 0.93
Difficulty range: -1.96 to 1.87

I.3 ELTS Subtests: Observed ICCs & Departures from Expectation

(i) G1 (Reading)

	ITEM CHARACTERISTIC CURVE						DEPARTURE FROM EXPECTED ICC					
ITEM NAME	SUBGROUP 1	2	3	4	5	6	SUBGROUP 1	2	3	4	5	6
G101	0.80	0.92	0.95	0.97	0.99	1.00	-0.02	0.01	0.01	0.01	0.02	0.00
G102	0.76	0.91	0.95	0.96	0.98	0.99	-0.03	0.02	0.02	0.01	0.01	0.00
G103	0.58	0.73	0.85	0.82	0.89	0.95	0.04	0.01	0.05	-0.03	-0.02	-0.01
G104	0.72	0.80	0.92	0.91	0.98	0.99	0.02	-0.04	0.03	-0.02	0.02	0.01
G105	0.60	0.82	0.82	0.85	0.92	0.97	0.02	0.06	-0.01	-0.03	-0.01	-0.01
G106	0.72	0.82	0.88	0.89	0.92	0.97	0.06	0.01	0.01	-0.01	-0.03	-0.01
G107	0.68	0.79	0.78	0.83	0.92	0.95	0.10	0.04	-0.04	-0.05	-0.01	-0.03
G108	0.64	0.87	0.90	0.88	0.93	0.96	-0.02	0.07	0.03	-0.02	-0.02	-0.02
G109	0.36	0.45	0.55	0.57	0.69	0.84	0.11	0.04	0.03	-0.05	-0.06	-0.06
G110	0.23	0.46	0.63	0.63	0.78	0.91	-0.05	0.01	0.07	-0.03	-0.00	0.01
G111	0.24	0.30	0.40	0.43	0.46	0.57	0.12	0.08	0.10	0.03	-0.10	-0.21
G112	0.25	0.34	0.42	0.50	0.63	0.77	0.07	0.03	0.01	-0.01	-0.02	-0.07
G113	0.64	0.80	0.90	0.92	0.99	1.00	-0.04	-0.02	0.02	0.01	0.04	0.02
G114	0.45	0.65	0.77	0.78	0.90	1.00	-0.02	-0.01	0.03	-0.03	0.01	0.04
G115	0.45	0.69	0.70	0.83	0.88	0.93	-0.00	0.05	-0.03	0.03	0.00	-0.02
G116	0.49	0.69	0.74	0.82	0.89	0.95	0.02	0.03	-0.01	-0.00	-0.00	-0.00
G117	0.21	0.42	0.56	0.67	0.74	0.88	-0.04	0.01	0.04	0.05	-0.01	-0.01
G118	0.38	0.56	0.66	0.72	0.82	0.95	0.02	0.01	0.01	-0.02	-0.01	0.01
G119	0.76	0.87	0.94	0.94	0.95	0.98	0.01	0.01	0.03	0.00	-0.02	-0.01
G120	0.67	0.87	0.92	0.91	0.93	0.99	-0.03	0.04	0.04	-0.01	-0.02	0.01
G121	0.47	0.66	0.75	0.87	0.93	0.98	-0.02	-0.02	-0.01	0.04	0.03	0.02
G122	0.57	0.76	0.80	0.88	0.96	0.99	-0.02	0.00	-0.02	0.01	0.03	0.02
G123	0.43	0.68	0.75	0.85	0.92	0.98	-0.05	0.01	-0.01	0.03	0.02	0.02
G124	0.36	0.58	0.61	0.72	0.81	0.90	0.02	0.06	-0.02	0.01	-0.02	-0.03
G125	0.22	0.28	0.37	0.42	0.51	0.69	0.09	0.04	0.04	-0.00	-0.06	-0.10
G126	0.14	0.24	0.33	0.44	0.60	0.82	0.00	-0.01	-0.01	0.00	0.01	0.02
G127	0.60	0.83	0.87	0.95	0.95	0.99	-0.05	0.02	0.01	0.04	0.00	0.01
G128	0.40	0.64	0.64	0.74	0.82	0.89	0.02	0.09	-0.02	-0.00	-0.02	-0.04
G129	0.46	0.65	0.75	0.83	0.88	0.98	-0.01	0.00	0.01	0.02	-0.01	0.02
G130	0.41	0.53	0.71	0.75	0.84	0.95	0.03	-0.04	0.04	-0.00	-0.01	0.01
G131	0.31	0.45	0.54	0.67	0.80	0.92	0.02	-0.01	-0.02	0.00	0.01	0.01
G132	0.27	0.42	0.50	0.66	0.73	0.86	0.02	0.02	-0.01	0.04	-0.02	-0.03
G133	0.42	0.67	0.82	0.93	0.98	0.99	-0.11	-0.05	0.02	0.08	0.06	0.03
G134	0.36	0.52	0.70	0.83	0.93	0.99	-0.06	-0.08	0.00	0.05	0.06	0.05
G135	0.15	0.25	0.37	0.49	0.71	0.92	-0.02	-0.06	-0.04	-0.02	0.06	0.08
G136	0.33	0.54	0.64	0.80	0.93	0.99	-0.06	-0.04	-0.04	0.04	0.08	0.05
G137	0.12	0.28	0.38	0.60	0.81	0.98	-0.10	-0.08	-0.08	0.04	0.10	0.11
G138	0.10	0.13	0.20	0.36	0.59	0.93	-0.02	-0.08	-0.09	-0.03	0.05	0.16
G139	0.30	0.45	0.58	0.69	0.84	0.94	-0.01	-0.03	-0.01	0.01	0.04	0.03
G140	0.21	0.32	0.50	0.70	0.83	0.99	-0.06	-0.11	-0.04	0.06	0.07	0.09

GROUP	SCORE RANGE	MEAN ABILITY	NO. IN SUBGROUP
1	1 - 21	-0.34	240
2	22 - 25	0.43	248
3	26 - 28	0.86	235
4	29 - 31	1.27	206
5	32 - 35	1.90	276
6	36 - 39	2.99	243

N =1448

(ii) G1 (Listening)

	ITEM CHARACTERISTIC CURVE						DEPARTURE FROM EXPECTED ICC					
ITEM NAME	SUBGROUP 1	2	3	4	5	6	SUBGROUP 1	2	3	4	5	6
G201	0.82	0.91	0.94	0.96	0.96	0.99	0.00	0.01	0.01	0.01	-0.00	0.00
G202	0.45	0.60	0.65	0.78	0.79	0.94	0.00	-0.01	-0.04	0.03	-0.02	0.04
G203	0.36	0.46	0.51	0.57	0.68	0.82	0.06	0.01	-0.02	-0.04	0.00	-0.01
G204	0.31	0.42	0.38	0.57	0.68	0.87	0.03	-0.00	-0.13	-0.01	0.01	0.06
G205	0.59	0.80	0.87	0.87	0.88	0.97	-0.04	0.03	0.04	0.01	-0.02	0.02
G206	0.43	0.58	0.74	0.76	0.84	0.95	-0.03	-0.05	0.04	-0.00	0.02	0.04
G207	0.41	0.64	0.68	0.74	0.80	0.90	-0.02	0.05	0.01	-0.00	0.00	0.00
G208	0.41	0.62	0.70	0.78	0.79	0.94	-0.04	0.01	0.01	0.02	-0.02	0.04
G209	0.32	0.52	0.58	0.65	0.76	0.90	-0.03	0.01	-0.02	-0.02	0.02	0.04
G210	0.47	0.61	0.65	0.85	0.86	0.91	-0.00	-0.02	-0.06	0.08	0.04	-0.00
G211	0.37	0.55	0.71	0.66	0.76	0.88	-0.01	0.00	0.08	-0.03	-0.01	0.00
G212	0.51	0.70	0.84	0.84	0.90	0.97	-0.06	-0.02	0.05	0.00	0.03	0.03
G213	0.18	0.25	0.35	0.40	0.47	0.69	0.02	-0.01	0.02	-0.00	-0.02	-0.00
G214	0.23	0.38	0.50	0.55	0.69	0.84	-0.04	-0.02	0.01	-0.02	0.04	0.04
G215	0.56	0.74	0.78	0.85	0.88	0.96	-0.02	0.01	-0.01	0.01	0.01	0.02
G216	0.23	0.38	0.43	0.56	0.67	0.84	-0.03	-0.02	-0.05	0.00	0.03	0.04
G217	0.66	0.85	0.86	0.91	0.92	0.98	-0.02	0.03	-0.01	0.01	-0.00	0.02
G218	0.69	0.88	0.93	0.99	0.97	0.99	-0.08	0.00	0.02	0.05	0.02	0.02
G219	0.69	0.84	0.88	0.88	0.92	0.98	-0.00	0.02	0.01	-0.01	-0.01	0.01
G220	0.82	0.94	0.98	0.98	0.99	0.99	-0.05	0.01	0.04	0.01	0.01	0.00
G221	0.57	0.77	0.81	0.84	0.89	0.97	-0.02	0.02	0.01	-0.01	0.01	0.02
G222	0.33	0.59	0.58	0.75	0.84	0.95	-0.08	0.02	-0.08	0.03	0.05	0.06
G223	0.44	0.64	0.72	0.80	0.88	0.96	-0.05	-0.02	-0.01	0.01	0.04	0.04
G224	0.58	0.77	0.84	0.88	0.93	0.97	-0.04	-0.00	0.01	0.02	0.03	0.02
G225	0.26	0.38	0.57	0.57	0.55	0.80	0.00	-0.01	0.10	0.02	-0.09	0.00
G226	0.69	0.75	0.86	0.83	0.88	0.94	0.07	-0.01	0.04	-0.03	-0.02	-0.01
G227	0.15	0.09	0.07	0.09	0.10	0.07	0.13	0.05	0.02	0.01	0.00	-0.16
G228	0.30	0.41	0.49	0.61	0.66	0.84	0.02	-0.02	-0.03	0.02	-0.01	0.02
G229	0.34	0.46	0.55	0.60	0.71	0.85	0.02	-0.01	-0.00	-0.02	0.01	0.01
G230	0.10	0.18	0.20	0.25	0.33	0.59	0.01	0.01	-0.02	-0.03	-0.02	0.02
G231	0.69	0.72	0.84	0.90	0.94	0.95	0.05	-0.06	-0.00	0.03	0.03	-0.01
G232	0.43	0.53	0.65	0.67	0.80	0.87	0.04	-0.02	0.02	-0.04	0.04	-0.01
G233	0.36	0.44	0.55	0.62	0.69	0.77	0.07	-0.00	0.02	0.01	0.00	-0.06
G234	0.34	0.50	0.54	0.62	0.71	0.83	0.03	0.03	-0.01	-0.01	0.00	-0.01
G235	0.14	0.22	0.29	0.33	0.35	0.42	0.05	0.05	0.08	0.05	-0.00	-0.14

GROUP	SCORE RANGE	MEAN ABILITY	NO. IN SUBGROUP
1	1 - 18	-0.35	259
2	19 - 21	0.33	227
3	22 - 13	0.67	195
4	24 - 25	0.98	207
5	26 - 27	1.32	225
6	28 - 34	2.24	368

N =1481

(iii) M1 (General Academic)

	ITEM CHARACTERISTIC CURVE						DEPARTURE FROM EXPECTED ICC					
ITEM NAME	SUBGROUP 1	2	3	4	5	6	SUBGROUP 1	2	3	4	5	6
GA01	0.62	0.67	0.73	0.83	0.88	0.88	0.12	0.02	-0.02	0.01	-0.02	-0.09
GA02	0.25	0.30	0.47	0.46	0.64	0.81	0.06	-0.00	0.06	-0.06	-0.03	-0.05
GA03	0.29	0.40	0.46	0.48	0.31	0.26	0.17	0.20	0.18	0.10	-0.23	-0.51
GA04	0.29	0.47	0.60	0.66	0.82	0.93	-0.03	0.01	0.02	-0.02	0.02	0.01
GA05	0.38	0.47	0.51	0.49	0.61	0.63	0.16	0.14	0.07	-0.07	-0.10	-0.24
GA06	0.43	0.53	0.59	0.77	0.79	0.89	0.07	0.02	-0.04	0.05	-0.04	-0.04
GA07	0.38	0.70	0.73	0.88	0.93	1.00	-0.12	0.05	-0.02	0.05	0.03	0.04
GA08	0.44	0.70	0.80	0.74	0.76	0.96	-0.01	0.09	0.09	-0.06	-0.12	0.01
GA09	0.13	0.26	0.39	0.37	0.64	0.82	-0.02	0.01	0.04	-0.08	0.03	0.01
GA10	0.41	0.51	0.71	0.71	0.85	0.93	0.02	-0.03	0.06	-0.04	0.00	-0.01
GA11	0.65	0.79	0.84	0.88	0.93	0.98	0.02	0.02	0.00	-0.01	-0.01	0.00
GA12	0.56	0.67	0.74	0.83	0.84	1.00	0.05	0.01	-0.01	0.00	-0.07	0.04
GA13	0.24	0.29	0.36	0.58	0.61	0.68	0.07	0.01	-0.02	0.10	-0.03	-0.15
GA14	0.24	0.31	0.47	0.68	0.70	0.95	-0.01	-0.06	-0.02	0.08	-0.04	0.06
GA15	0.38	0.43	0.51	0.65	0.75	0.91	0.09	-0.00	-0.03	-0.01	-0.04	0.00
GA16	0.25	0.46	0.39	0.60	0.63	0.86	0.03	0.11	-0.07	0.03	-0.09	-0.02
GA17	0.43	0.57	0.61	0.69	0.69	0.88	0.09	0.09	0.02	-0.01	-0.13	-0.05
GA18	0.11	0.24	0.21	0.45	0.48	0.82	-0.01	0.04	-0.07	0.06	-0.07	0.05
GA19	0.24	0.41	0.37	0.55	0.64	0.82	0.03	0.09	-0.06	0.01	-0.05	-0.04
GA20	0.16	0.40	0.51	0.72	0.90	1.00	-0.14	-0.04	-0.04	0.06	0.10	0.08
GA21	0.19	0.17	0.23	0.38	0.51	0.91	0.06	-0.04	-0.07	-0.01	-0.05	0.13
GA22	0.40	0.67	0.74	0.83	0.91	1.00	-0.09	0.03	0.00	0.01	0.02	0.04
GA23	0.24	0.43	0.56	0.72	0.84	0.96	-0.07	-0.03	-0.01	0.05	0.04	0.05
GA24	0.14	0.24	0.56	0.71	0.87	0.98	-0.13	-0.16	0.04	0.08	0.10	0.08
GA25	0.10	0.19	0.27	0.35	0.48	0.84	-0.02	-0.00	0.00	-0.01	-0.05	0.08
GA26	0.14	0.17	0.17	0.28	0.52	0.47	0.06	0.04	-0.03	-0.00	0.10	-0.21
GA27	0.27	0.43	0.51	0.62	0.75	0.86	0.01	0.04	0.01	-0.00	-0.01	-0.04
GA28	0.30	0.53	0.57	0.74	0.88	1.00	-0.06	0.02	-0.06	0.01	0.05	0.07
GA29	0.22	0.26	0.34	0.52	0.69	0.91	0.03	-0.04	-0.06	0.01	0.02	0.06
GA30	0.10	0.33	0.29	0.35	0.52	0.68	-0.02	0.13	0.00	-0.03	-0.02	-0.09
GA31	0.21	0.36	0.47	0.54	0.82	0.96	-0.04	-0.02	-0.02	-0.06	0.08	0.07
GA32	0.11	0.40	0.61	0.62	0.78	0.91	-0.15	0.01	0.11	-0.00	0.02	0.01
GA33	0.13	0.06	0.31	0.37	0.61	0.86	0.00	-0.15	0.03	-0.02	0.06	0.08
GA34	0.13	0.10	0.34	0.54	0.81	0.98	-0.05	-0.19	-0.05	0.03	0.15	0.14
GA35	0.08	0.14	0.17	0.23	0.43	0.75	-0.00	0.01	-0.03	-0.05	0.00	0.06
GA36	0.06	0.16	0.51	0.51	0.81	0.91	-0.13	-0.15	0.10	-0.01	0.13	0.06
GA37	0.05	0.07	0.27	0.37	0.75	0.95	-0.09	-0.15	-0.04	-0.04	0.17	0.15
GA38	0.11	0.16	0.30	0.42	0.70	0.84	-0.03	-0.07	-0.02	-0.02	0.11	0.04
GA39	0.02	0.06	0.13	0.23	0.48	0.86	-0.06	-0.07	-0.06	-0.04	0.06	0.18
GA40	0.05	0.07	0.27	0.40	0.66	0.96	-0.08	-0.14	-0.03	-0.00	0.09	0.18

GROUP	SCORE RANGE	MEAN ABILITY	NO. IN SUBGROUP
1	1 - 12	-1.24	63
2	13 - 16	-0.61	70
3	17 - 20	-0.14	70
4	21 - 25	0.32	65
5	26 - 31	0.97	67
6	32 - 39	2.11	57

N = 392

(iv) M1 (Life Sciences)

	ITEM CHARACTERISTIC CURVE						DEPARTURE FROM EXPECTED ICC					
ITEM NAME	SUBGROUP 1	2	3	4	5	6	SUBGROUP 1	2	3	4	5	6
LS01	0.74	0.78	0.81	0.86	0.85	0.93	0.11	0.03	-0.00	-0.01	-0.05	-0.03
LS02	0.79	0.89	0.89	0.93	0.97	0.97	0.02	0.02	-0.02	-0.01	0.01	-0.01
LS03	0.47	0.63	0.67	0.80	0.79	0.88	0.02	0.04	-0.01	0.04	-0.04	-0.03
LS04	0.38	0.51	0.50	0.52	0.59	0.51	0.15	0.17	0.07	-0.01	-0.04	-0.27
LS05	0.89	0.97	0.96	0.98	0.98	1.00	-0.02	0.02	-0.00	0.01	0.00	0.01
LS06	0.89	0.94	0.96	0.98	0.98	0.94	0.02	0.01	0.02	0.02	0.01	-0.05
LS07	0.64	0.73	0.78	0.86	0.90	1.00	0.03	-0.01	-0.03	-0.01	-0.00	0.05
LS08	0.30	0.16	0.35	0.34	0.34	0.44	0.19	-0.02	0.11	0.02	-0.07	-0.17
LS09	0.72	0.79	0.89	0.86	0.89	0.96	0.05	0.01	0.05	-0.03	-0.04	-0.01
LS10	0.40	0.60	0.70	0.68	0.69	0.84	0.02	0.09	0.09	-0.02	-0.09	-0.05
LS11	0.57	0.65	0.72	0.73	0.79	0.85	0.11	0.05	0.03	-0.04	-0.04	-0.06
LS12	0.66	0.71	0.83	0.80	0.92	0.91	0.07	-0.01	0.04	-0.05	0.02	-0.04
LS13	0.62	0.78	0.83	0.86	0.92	0.96	-0.01	0.02	0.01	-0.01	0.01	0.00
LS14	0.28	0.43	0.39	0.34	0.54	0.71	0.09	0.13	0.01	-0.14	-0.04	-0.04
LS15	0.66	0.62	0.83	0.89	0.85	0.94	0.08	-0.09	0.05	0.05	-0.04	-0.00
LS16	0.38	0.52	0.57	0.57	0.69	0.65	0.09	0.11	0.07	-0.03	-0.00	-0.18
LS17	0.32	0.49	0.46	0.62	0.74	0.82	0.02	0.06	-0.07	-0.00	0.03	-0.02
LS18	0.62	0.79	0.80	0.84	0.90	0.90	0.03	0.07	-0.00	-0.01	0.01	-0.05
LS19	0.51	0.71	0.70	0.79	0.79	0.87	0.03	0.10	-0.00	0.01	-0.05	-0.05
LS20	0.40	0.57	0.72	0.75	0.82	0.81	-0.01	0.02	0.08	0.02	0.02	-0.09
LS21	0.49	0.63	0.78	0.82	0.85	0.88	-0.01	-0.00	0.06	0.03	0.00	-0.04
LS22	0.17	0.37	0.41	0.46	0.41	0.65	0.00	0.11	0.07	0.03	-0.12	-0.07
LS23	0.58	0.86	0.89	0.95	0.95	0.97	-0.12	0.05	0.02	0.04	0.02	0.00
LS24	0.58	0.78	0.89	0.95	0.95	0.99	-0.10	-0.02	0.04	0.05	0.02	0.02
LS25	0.38	0.57	0.76	0.71	0.79	0.91	-0.05	0.01	0.10	-0.03	-0.02	0.01
LS26	0.26	0.33	0.54	0.54	0.61	0.84	0.01	-0.03	0.07	-0.02	-0.05	0.03
LS27	0.09	0.30	0.44	0.52	0.46	0.91	-0.11	-0.00	0.05	0.02	-0.13	0.15
LS28	0.19	0.25	0.39	0.43	0.59	0.84	-0.01	-0.04	0.00	-0.05	0.01	0.09
LS29	0.23	0.41	0.56	0.77	0.89	0.99	-0.15	-0.10	-0.05	0.07	0.11	0.10
LS30	0.17	0.29	0.57	0.71	0.74	0.82	-0.11	-0.12	0.08	0.12	0.05	-0.00
LS31	0.09	0.24	0.31	0.66	0.74	0.90	-0.14	-0.11	-0.12	0.13	0.11	0.11
LS32	0.09	0.21	0.28	0.50	0.59	0.84	-0.08	-0.07	-0.08	0.05	0.04	0.11
LS33	0.25	0.25	0.33	0.55	0.75	0.93	-0.00	-0.11	-0.13	-0.00	0.10	0.12
LS34	0.08	0.16	0.22	0.30	0.48	0.75	-0.05	-0.04	-0.04	-0.05	0.03	0.11
LS35	0.04	0.02	0.02	0.05	0.07	0.04	0.03	0.00	-0.00	0.02	0.02	-0.07
LS36	0.15	0.22	0.28	0.32	0.56	0.65	0.01	0.00	-0.02	-0.06	0.08	-0.02
LS37	0.08	0.17	0.30	0.46	0.59	0.88	-0.10	-0.09	-0.06	0.02	0.05	0.16
LS38	0.09	0.17	0.17	0.29	0.52	0.85	-0.04	-0.04	-0.12	-0.09	0.05	0.19
LS39	0.11	0.21	0.31	0.55	0.69	0.93	-0.10	-0.11	-0.10	0.04	0.08	0.16
LS40	0.11	0.16	0.20	0.50	0.66	0.90	-0.07	-0.12	-0.16	0.04	0.10	0.16

GROUP	SCORE RANGE	MEAN ABILITY	NO. IN SUBGROUP
1	1 - 18	-0.59	53
2	19 - 21	-0.02	63
3	22 - 24	0.38	54
4	25 - 27	0.77	56
5	28 - 30	1.16	61
6	31 - 39	2.01	68

N = 355

(v) M1 (Medicine)

	ITEM CHARACTERISTIC CURVE						DEPARTURE FROM EXPECTED ICC					
ITEM NAME	SUBGROUP 1	2	3	4	5	6	SUBGROUP 1	2	3	4	5	6
ME01	0.71	0.76	0.96	0.96	0.96	1.00	0.00	-0.09	0.05	0.03	0.00	0.02
ME02	0.71	0.90	0.96	1.00	1.00	0.95	-0.06	0.01	0.03	0.05	0.03	-0.03
ME03	0.46	0.71	0.87	0.92	0.96	0.95	-0.09	-0.03	0.05	0.05	0.04	-0.01
ME04	0.67	0.71	0.74	0.81	0.78	0.95	0.19	0.04	-0.03	-0.03	-0.12	0.01
ME05	0.54	0.48	0.83	0.81	0.70	0.76	0.18	-0.08	0.16	0.05	-0.13	-0.15
ME06	0.17	0.33	0.30	0.35	0.37	0.71	0.06	0.13	0.01	-0.05	-0.14	0.04
ME07	0.71	0.86	0.83	0.85	0.96	0.95	0.08	0.06	-0.04	-0.06	0.02	-0.02
ME08	0.71	0.86	0.96	0.81	1.00	0.90	0.05	0.03	0.08	-0.11	0.05	-0.07
ME09	0.54	0.90	0.87	1.00	0.93	0.95	-0.10	0.09	-0.00	0.09	-0.02	-0.02
ME10	0.54	0.62	0.83	1.00	0.93	0.95	-0.01	-0.12	0.01	0.12	0.01	-0.01
ME11	0.79	0.90	0.96	1.00	0.93	0.95	0.02	0.01	0.03	0.05	-0.05	-0.03
ME12	0.83	0.86	1.00	0.92	0.96	1.00	0.05	-0.05	0.06	-0.04	-0.01	0.01
ME13	0.08	0.24	0.22	0.31	0.48	0.43	0.00	0.08	-0.00	-0.00	0.06	-0.17
ME14	0.58	0.81	0.91	0.77	0.89	0.95	0.03	0.07	0.10	-0.11	-0.03	-0.01
ME15	0.46	0.76	0.74	0.81	0.93	0.90	-0.01	0.09	-0.02	-0.03	0.04	-0.04
ME16	0.67	0.76	0.87	0.96	0.89	1.00	0.04	-0.04	0.00	0.05	-0.05	0.03
ME17	0.75	0.95	1.00	1.00	1.00	1.00	-0.09	0.02	0.04	0.03	0.02	0.01
ME18	0.79	0.90	0.96	1.00	0.93	1.00	0.00	-0.00	0.02	0.04	-0.05	0.01
ME19	0.54	0.71	0.96	0.81	1.00	1.00	-0.05	-0.06	0.11	-0.09	0.07	0.04
ME20	0.75	0.71	0.87	0.73	0.89	1.00	0.19	-0.04	0.04	-0.15	-0.03	0.04
ME21	0.17	0.29	0.52	0.62	0.56	0.76	-0.01	-0.03	0.10	0.08	-0.09	-0.03
ME22	0.71	0.81	0.87	0.81	0.89	1.00	0.11	0.03	0.02	-0.09	-0.05	0.03
ME23	0.25	0.48	0.61	0.69	0.93	1.00	-0.08	-0.05	-0.02	-0.04	0.11	0.10
ME24	0.21	0.38	0.57	0.62	0.74	0.81	-0.02	-0.01	0.06	-0.00	0.02	-0.03
ME25	0.37	0.24	0.61	0.62	0.70	0.86	0.13	-0.17	0.09	-0.02	-0.03	0.01
ME26	0.33	0.33	0.48	0.42	0.78	0.86	0.12	-0.04	-0.00	-0.17	0.08	0.03
ME27	0.25	0.29	0.74	0.62	0.85	0.95	-0.04	-0.18	0.16	-0.07	0.07	0.08
ME28	0.17	0.33	0.52	0.58	0.67	0.90	-0.04	-0.03	0.05	-0.01	-0.03	0.09
ME29	0.37	0.76	0.65	0.81	0.96	1.00	-0.09	0.10	-0.10	-0.02	0.08	0.06
ME30	0.21	0.38	0.30	0.73	0.81	0.81	-0.02	-0.00	-0.19	0.12	0.10	-0.02
ME31	0.37	0.71	0.91	0.96	0.96	1.00	-0.19	-0.04	0.09	0.08	0.04	0.04
ME32	0.29	0.48	0.57	0.54	0.74	0.86	0.04	0.06	0.03	-0.10	-0.00	0.01
ME33	0.62	0.95	0.83	1.00	1.00	1.00	-0.10	0.09	-0.08	0.06	0.04	0.02
ME34	0.17	0.62	0.65	0.88	0.89	1.00	-0.22	0.04	-0.03	0.11	0.04	0.08
ME35	0.17	0.38	0.43	0.42	0.37	0.57	0.04	0.16	0.12	0.01	-0.17	-0.13
ME36	0.21	0.43	0.43	0.81	0.81	0.95	-0.07	-0.03	-0.14	0.13	0.04	0.08
ME37	0.46	0.62	0.52	0.92	0.89	0.95	0.04	0.00	-0.20	0.12	0.02	0.02
ME38	0.42	0.57	0.52	0.92	0.93	1.00	-0.00	-0.05	-0.20	0.12	0.06	0.07
ME39	0.12	0.43	0.22	0.62	0.78	0.90	-0.07	0.08	-0.24	0.05	0.10	0.10
ME40	0.21	0.43	0.48	0.58	0.81	0.76	-0.02	0.04	-0.02	-0.03	0.10	-0.07

GROUP	SCORE RANGE	MEAN ABILITY	NO. IN SUBGROUP
1	1 - 22	-0.27	24
2	23 - 26	0.58	21
3	27 - 29	1.04	23
4	30 - 32	1.49	26
5	33 - 34	1.97	27
6	35 - 39	2.71	21

N = 142

(vi) M1 (Physical Sciences)

	ITEM CHARACTERISTIC CURVE						DEPARTURE FROM EXPECTED ICC					
ITEM NAME	SUBGROUP 1	2	3	4	5	6	SUBGROUP 1	2	3	4	5	6
PS01	0.80	1.00	0.89	1.00	0.95	0.97	0.02	0.09	-0.06	0.03	-0.03	-0.02
PS02	0.35	0.57	0.56	0.84	0.95	0.97	0.01	-0.01	-0.16	0.04	0.09	0.03
PS03	0.55	0.78	0.94	0.95	0.95	1.00	-0.05	-0.02	0.06	0.02	-0.00	0.02
PS04	0.10	0.09	0.33	0.47	0.40	0.69	0.03	-0.08	0.06	0.11	-0.08	-0.01
PS05	0.50	0.65	0.89	0.89	0.90	1.00	0.01	-0.07	0.06	0.01	-0.02	0.03
PS06	0.30	0.57	0.61	0.63	0.65	0.78	0.09	0.16	0.05	-0.03	-0.10	-0.10
PS07	0.10	0.30	0.50	0.47	0.65	0.78	-0.03	0.03	0.08	-0.05	0.02	-0.03
PS08	0.70	0.96	0.94	0.95	0.90	0.97	0.00	0.09	0.02	-0.00	-0.07	-0.02
PS09	0.80	0.96	1.00	1.00	1.00	1.00	-0.06	0.01	0.03	0.02	0.01	0.00
PS10	0.40	0.74	0.78	0.79	1.00	0.97	-0.05	0.06	-0.03	-0.07	0.09	0.01
PS11	0.75	0.96	0.94	1.00	1.00	0.97	-0.03	0.04	-0.01	0.03	0.02	-0.02
PS12	0.35	0.48	0.83	0.74	0.85	0.84	0.05	-0.04	0.16	-0.02	0.02	-0.08
PS13	0.15	0.22	0.61	0.53	0.85	0.94	-0.04	-0.15	0.09	-0.10	0.13	0.07
PS14	0.50	0.48	0.78	0.74	0.80	1.00	0.15	-0.11	0.05	-0.07	-0.07	0.06
PS15	0.40	0.74	0.78	0.68	0.75	0.94	0.05	0.15	0.05	-0.12	-0.12	-0.00
PS16	0.70	0.96	0.94	0.95	1.00	1.00	-0.06	0.05	-0.00	-0.02	0.02	0.01
PS17	0.60	0.74	1.00	0.84	0.95	0.94	0.05	-0.03	0.14	-0.07	0.01	-0.04
PS18	0.65	0.65	0.78	0.89	0.90	0.97	0.16	-0.07	-0.05	0.01	-0.02	0.00
PS19	0.70	0.87	0.89	1.00	1.00	1.00	-0.02	-0.01	-0.04	0.04	0.03	0.01
PS20	0.65	0.83	0.89	0.95	1.00	0.97	0.01	-0.00	-0.01	0.01	0.04	-0.01
PS21	0.35	0.65	0.72	0.89	0.90	1.00	-0.06	0.01	-0.05	0.05	0.01	0.05
PS22	0.85	0.74	0.83	1.00	1.00	1.00	0.15	-0.13	-0.09	0.05	0.03	0.01
PS23	0.55	0.78	0.78	1.00	0.95	1.00	-0.02	0.00	-0.09	0.09	0.01	0.02
PS24	0.70	0.61	0.83	0.84	1.00	0.94	0.20	-0.12	-0.01	-0.05	0.07	-0.03
PS25	0.55	0.78	0.72	0.79	1.00	0.97	0.06	0.06	-0.11	-0.09	0.08	0.00
PS26	0.35	0.61	0.83	0.95	0.85	1.00	-0.07	-0.05	0.05	0.10	-0.05	0.05
PS27	0.55	0.91	0.83	0.95	1.00	1.00	-0.09	0.08	-0.07	0.01	0.04	0.02
PS28	0.30	0.74	0.94	0.79	0.90	0.94	-0.12	0.08	0.16	-0.06	0.00	-0.02
PS29	0.20	0.39	0.56	0.68	0.90	0.97	-0.05	-0.06	-0.06	-0.02	0.11	0.07
PS30	0.50	0.91	0.83	0.95	1.00	1.00	-0.12	0.09	-0.06	0.02	0.05	0.02
PS31	0.25	0.39	0.39	0.74	0.85	0.97	0.02	-0.04	-0.20	0.05	0.07	0.08
PS32	0.35	0.57	0.72	0.84	0.55	0.91	0.07	0.07	0.07	0.10	-0.27	-0.01
PS33	0.30	0.57	0.78	0.79	0.90	0.97	-0.05	-0.02	0.05	-0.02	0.03	0.03
PS34	0.40	0.91	0.89	0.95	1.00	1.00	-0.20	0.11	0.00	0.02	0.05	0.02
PS35	0.20	0.39	0.56	0.63	0.60	0.84	0.03	0.04	0.05	0.02	-0.11	-0.01
PS36	0.40	0.57	0.67	0.79	0.95	1.00	0.02	-0.04	-0.08	-0.03	0.07	0.06
PS37	0.30	0.35	0.78	0.84	1.00	0.94	-0.02	-0.20	0.08	0.06	0.15	0.01
PS38	0.15	0.61	0.61	0.79	0.85	1.00	-0.15	0.08	-0.07	0.02	0.01	0.07
PS39	0.30	0.35	0.61	0.74	0.60	0.91	0.09	-0.06	0.05	0.07	-0.15	0.02
PS40	0.00	0.17	0.22	0.16	0.20	0.12	-0.02	0.13	0.15	0.05	0.03	-0.22

GROUP	SCORE RANGE	MEAN ABILITY	NO. IN SUBGROUP
1	1 - 22	-0.35	20
2	23 - 27	0.66	23
3	28 - 31	1.31	18
4	32 - 33	1.73	19
5	34 - 35	2.18	20
6	36 - 39	3.15	32

N = 132

(vii) M1 (Social Studies)

	ITEM CHARACTERISTIC CURVE						DEPARTURE FROM EXPECTED ICC					
ITEM NAME	SUBGROUP 1	2	3	4	5	6	SUBGROUP 1	2	3	4	5	6
SS01	0.18	0.37	0.40	0.49	0.77	0.78	-0.04	0.02	-0.04	-0.04	0.14	-0.00
SS02	0.35	0.47	0.57	0.67	0.85	0.89	-0.01	-0.03	-0.02	-0.01	0.09	0.02
SS03	0.22	0.57	0.60	0.64	0.69	0.89	-0.10	0.10	0.03	-0.00	-0.04	0.04
SS04	0.53	0.70	0.68	0.64	0.74	0.86	0.10	0.11	0.01	-0.10	-0.07	-0.04
SS05	0.08	0.25	0.21	0.22	0.49	0.73	-0.03	0.05	-0.05	-0.11	0.06	0.10
SS06	0.45	0.55	0.47	0.76	0.62	0.89	0.10	0.05	-0.12	0.08	-0.14	0.02
SS07	0.39	0.62	0.62	0.71	0.87	0.89	-0.02	0.05	-0.04	-0.02	0.07	-0.01
SS08	0.18	0.25	0.32	0.33	0.49	0.81	0.03	-0.01	-0.01	-0.08	-0.03	0.11
SS09	0.59	0.72	0.81	0.82	0.95	1.00	-0.01	-0.02	0.00	-0.03	0.05	0.05
SS10	0.43	0.42	0.57	0.62	0.67	0.92	0.10	-0.06	0.00	-0.03	-0.07	0.06
SS11	0.20	0.40	0.43	0.71	0.74	0.97	-0.09	-0.04	-0.10	0.10	0.04	0.13
SS12	0.12	0.17	0.34	0.51	0.56	0.73	-0.04	-0.09	-0.01	0.08	0.04	0.02
SS13	0.45	0.55	0.53	0.82	0.85	1.00	0.02	-0.04	-0.14	0.08	0.03	0.10
SS14	0.43	0.55	0.53	0.60	0.67	0.68	0.13	0.10	-0.01	-0.03	-0.05	-0.17
SS15	0.63	0.72	0.85	0.84	0.85	0.97	0.04	-0.02	0.04	-0.01	-0.05	0.02
SS16	0.84	0.85	0.94	0.96	0.97	0.97	0.03	-0.04	0.01	0.01	0.01	-0.01
SS17	0.82	0.97	0.94	0.98	0.97	0.95	-0.02	0.06	-0.00	0.02	0.00	-0.04
SS18	0.63	0.82	0.91	0.98	0.97	1.00	-0.09	-0.02	0.03	0.06	0.03	0.03
SS19	0.10	0.17	0.11	0.27	0.21	0.05	0.06	0.09	-0.00	0.12	-0.00	-0.34
SS20	0.43	0.60	0.77	0.67	0.69	0.86	0.03	0.04	0.12	-0.06	-0.10	-0.03
SS21	0.24	0.37	0.40	0.49	0.59	0.76	0.03	0.04	-0.02	-0.02	-0.02	-0.02
SS22	0.31	0.37	0.38	0.40	0.38	0.65	0.13	0.09	0.02	-0.05	-0.16	-0.08
SS23	0.24	0.30	0.55	0.56	0.79	0.78	-0.02	-0.10	0.06	-0.02	0.12	-0.03
SS24	0.35	0.50	0.49	0.42	0.62	0.70	0.10	0.12	0.02	-0.13	-0.03	-0.10
SS25	0.37	0.47	0.57	0.62	0.72	0.78	0.05	0.01	0.02	-0.02	-0.01	-0.07
SS26	0.35	0.40	0.62	0.58	0.56	0.86	0.06	-0.03	0.09	-0.03	-0.13	0.03
SS27	0.37	0.55	0.81	0.82	0.87	0.92	-0.10	-0.07	0.10	0.05	0.04	0.00
SS28	0.24	0.50	0.57	0.78	0.82	0.97	-0.12	-0.02	-0.04	0.09	0.05	0.09
SS29	0.16	0.20	0.28	0.44	0.62	0.70	0.00	-0.06	-0.06	0.02	0.10	-0.00
SS30	0.08	0.27	0.38	0.42	0.72	0.81	-0.11	-0.03	-0.00	-0.05	0.15	0.07
SS31	0.20	0.30	0.64	0.53	0.69	0.97	-0.07	-0.12	0.13	-0.06	0.01	0.14
SS32	0.12	0.25	0.47	0.56	0.64	0.92	-0.10	-0.10	0.03	0.03	0.02	0.14
SS33	0.18	0.37	0.60	0.73	0.85	0.92	-0.14	-0.10	0.03	0.08	0.11	0.06
SS34	0.16	0.27	0.40	0.44	0.44	0.76	-0.01	-0.00	0.05	0.01	-0.10	0.04
SS35	0.22	0.47	0.53	0.62	0.64	0.89	-0.06	0.05	0.01	0.02	-0.05	0.06
SS36	0.20	0.17	0.30	0.33	0.33	0.49	0.09	-0.01	0.05	0.01	-0.08	-0.13
SS37	0.24	0.52	0.53	0.73	0.87	0.92	-0.11	0.02	-0.07	0.05	0.11	0.05
SS38	0.14	0.20	0.28	0.40	0.51	0.81	-0.01	-0.05	-0.05	-0.01	0.01	0.12
SS39	0.22	0.27	0.45	0.51	0.59	0.76	0.02	-0.05	0.03	0.01	-0.01	-0.01
SS40	0.10	0.25	0.15	0.40	0.36	0.41	0.01	0.09	-0.07	0.12	-0.01	-0.17

GROUP	SCORE RANGE	MEAN ABILITY	NO. IN SUBGROUP
1	1 - 15	-0.89	49
2	16 - 19	-0.22	40
3	20 - 22	0.14	47
4	23 - 25	0.50	45
5	26 - 29	0.89	39
6	30 - 39	1.77	37

N = 257

(viii) M1 (Technology)

	ITEM CHARACTERISTIC CURVE						DEPARTURE FROM EXPECTED ICC					
ITEM NAME	SUBGROUP 1	2	3	4	5	6	SUBGROUP 1	2	3	4	5	6
TN01	0.73	0.89	0.83	0.94	0.86	0.93	0.12	0.04	-0.06	0.02	-0.09	-0.04
TN02	0.52	0.77	0.87	0.83	0.93	0.96	0.01	0.00	0.04	-0.04	0.02	0.01
TN03	0.67	0.97	1.00	0.97	1.00	1.00	-0.10	0.04	0.05	0.01	0.02	0.01
TN04	0.64	0.91	0.96	1.00	1.00	1.00	-0.08	0.00	0.02	0.05	0.03	0.02
TN05	0.48	0.74	0.87	0.91	0.82	0.96	-0.00	-0.01	0.06	0.05	-0.08	0.02
TN06	0.88	0.94	0.87	1.00	0.96	1.00	0.07	-0.01	-0.09	0.03	-0.02	0.01
TN07	0.67	0.89	1.00	1.00	1.00	1.00	-0.07	-0.03	0.06	0.04	0.03	0.01
TN08	0.73	0.94	1.00	1.00	1.00	0.96	-0.05	0.01	0.04	0.03	0.02	-0.02
TN09	0.64	0.86	0.96	0.91	0.89	0.96	0.03	0.01	0.07	-0.01	-0.05	-0.01
TN10	0.39	0.46	0.78	0.77	0.96	1.00	0.02	-0.18	0.06	-0.01	0.12	0.09
TN11	0.36	0.54	0.57	0.71	0.79	0.89	0.07	-0.01	-0.07	0.00	0.00	0.01
TN12	0.24	0.51	0.65	0.80	0.71	0.82	-0.03	-0.01	0.04	0.11	-0.05	-0.05
TN13	0.45	0.46	0.43	0.49	0.43	0.61	0.28	0.10	-0.01	-0.05	-0.20	-0.17
TN14	0.52	0.86	0.87	0.91	0.96	0.96	-0.05	0.04	0.00	0.01	0.03	-0.00
TN15	0.58	0.91	0.96	1.00	1.00	1.00	-0.12	0.02	0.03	0.05	0.04	0.02
TN16	0.39	0.94	0.87	0.94	0.93	0.96	-0.17	0.13	0.01	0.04	-0.00	0.00
TN17	0.27	0.54	0.70	0.91	0.82	0.89	-0.07	-0.06	0.01	0.16	-0.00	-0.01
TN18	0.36	0.66	0.65	0.86	0.93	0.93	-0.03	-0.01	-0.09	0.06	0.07	0.01
TN19	0.42	0.71	0.70	0.80	0.93	0.96	0.00	0.02	-0.07	-0.02	0.06	0.03
TN20	0.39	0.80	0.83	0.77	0.93	0.96	-0.06	0.08	0.04	-0.07	0.04	0.02
TN21	0.24	0.46	0.43	0.77	0.57	0.82	0.02	0.02	-0.10	0.16	-0.13	-0.01
TN22	0.45	0.54	0.43	0.57	0.79	0.93	0.18	0.02	-0.17	-0.12	0.02	0.06
TN23	0.30	0.54	0.65	0.66	0.96	0.93	-0.02	-0.03	-0.01	-0.08	0.16	0.04
TN24	0.55	0.71	0.61	0.71	0.71	0.86	0.19	0.09	-0.09	-0.05	-0.12	-0.05
TN25	0.58	0.83	0.96	0.91	1.00	1.00	-0.05	-0.03	0.06	-0.01	0.05	0.03
TN26	0.48	0.74	0.87	0.86	0.86	1.00	-0.00	-0.01	0.06	-0.00	-0.04	0.05
TN27	0.61	0.66	0.61	0.63	0.68	0.96	0.26	0.04	-0.08	-0.13	-0.15	0.06
TN28	0.48	0.57	0.57	0.63	0.86	0.96	0.15	-0.02	-0.11	-0.12	0.04	0.07
TN29	0.52	0.74	0.83	0.86	0.89	1.00	0.02	-0.02	0.01	-0.01	-0.01	0.05
TN30	0.45	0.74	0.83	0.89	0.71	0.89	0.03	0.05	0.06	0.06	-0.16	-0.04
TN31	0.45	0.54	0.96	0.94	1.00	0.96	-0.03	-0.21	0.14	0.08	0.10	0.02
TN32	0.61	0.97	0.96	1.00	1.00	1.00	-0.13	0.05	0.01	0.04	0.03	0.01
TN33	0.24	0.51	0.70	0.60	0.86	0.86	-0.03	-0.00	0.09	-0.08	0.10	-0.01
TN34	0.36	0.91	0.96	1.00	1.00	0.93	-0.23	0.08	0.08	0.09	0.06	-0.04
TN35	0.18	0.51	0.39	0.54	0.82	0.86	-0.04	0.08	-0.13	-0.07	0.12	0.03
TN36	0.24	0.71	0.70	0.89	1.00	0.93	-0.17	0.04	-0.06	0.07	0.14	0.00
TN37	0.42	0.54	0.87	0.86	0.96	0.96	-0.01	-0.16	0.10	0.03	0.09	0.03
TN38	0.30	0.63	0.74	0.66	0.61	0.71	0.04	0.13	0.15	-0.01	-0.14	-0.14
TN39	0.15	0.29	0.43	0.40	0.61	0.68	0.02	-0.01	0.06	-0.06	0.05	-0.04
TN40	0.27	0.43	0.61	0.66	0.79	0.96	0.01	-0.08	0.01	-0.02	0.03	0.10

GROUP	SCORE RANGE	MEAN ABILITY	NO. IN SUBGROUP
1	1 - 25	-0.26	33
2	26 - 29	0.98	35
3	30 - 31	1.34	23
4	32 - 33	1.69	35
5	34 - 35	2.08	28
6	36 - 39	2.86	28

N = 182

I.4 ELTS Subtests: Item Fit Statistics

(i) G1 (Reading)

(Items ordered by total fit-t; 32 misfitting persons omitted)

ITEM NAME	ITEM DIFFIC.	BETWEEN GROUP FIT-T	TOTAL FIT-T	RASCH DISCRIM. INDEX
G137	1.02	6.82	-8.91	1.50
G138	1.73	6.32	-7.14	1.43
G140	0.71	5.73	-6.57	1.40
G135	1.25	3.41	-4.59	1.28
G134	0.02	4.22	-4.01	1.30
G136	0.12	4.25	-3.82	1.30
G133	-0.49	5.39	-3.81	1.35
G123	-0.26	1.41	-2.51	1.16
G139	0.49	0.43	-2.07	1.12
G121	-0.33	1.35	-1.77	1.14
G113	-1.11	2.58	-1.67	1.19
G127	-0.98	1.35	-1.57	1.11
G114	-0.20	1.92	-1.51	1.12
G122	-0.70	1.33	-1.32	1.12
G110	0.65	1.21	-1.29	1.06
G129	-0.20	-0.21	-1.10	1.06
G101	-1.90	0.23	-1.09	1.11
G102	-1.69	-0.13	-1.08	1.07
G120	-1.20	1.65	-0.88	0.99
G117	0.77	0.64	-0.84	1.02
G104	-1.24	1.41	-0.73	1.03
G126	1.52	-1.78	-0.69	1.05
G119	-1.44	0.86	-0.58	0.89
G115	-0.14	0.74	-0.23	0.94
G108	-0.99	2.82	-0.22	0.87
G130	0.16	0.22	0.02	0.99
G116	-0.22	-1.41	0.09	0.94
G131	0.59	-1.19	0.20	1.02
G118	0.25	-1.03	0.49	0.97
G105	-0.70	1.11	0.53	0.88
G103	-0.51	1.56	0.65	0.85
G106	-1.02	1.75	1.00	0.79
G124	0.35	0.93	1.11	0.86
G107	-0.66	3.56	1.71	0.72
G128	0.21	2.58	1.91	0.80
G132	0.82	0.46	2.05	0.89
G112	1.24	3.13	4.36	0.72
G109	0.79	4.50	5.45	0.64
G125	1.60	5.01	5.84	0.61
G111	1.69	9.22	9.56	0.32

TOTAL FIT-T:
Mean = -0.63
SD = 3.37
Range = -8.91 to 9.56

BETWEEN-GROUP FIT-T:
Mean = 2.11
SD = 2.41
Range = -1.78 to 9.22

DISCRIM. INDEX:
Range = 0.32 to 1.50

(ii) G2 (Listening)

(Items ordered by total fit-t; 22 misfitting persons omitted)

ITEM NAME	ITEM DIFFIC.	BETWEEN GROUP FIT-T	TOTAL FIT-T	RASCH DISCRIM. INDEX
G222	0.03	3.90	-3.59	1.30
G216	0.76	1.22	-2.85	1.18
G223	-0.33	2.55	-2.69	1.25
G206	-0.19	1.97	-2.62	1.20
G214	0.71	0.87	-2.19	1.17
G212	-0.64	2.17	-2.18	1.21
G204	0.65	3.22	-2.03	1.12
G218	-1.63	3.94	-2.00	1.35
G209	0.29	0.67	-1.87	1.14
G208	-0.14	1.33	-1.78	1.15
G224	-0.87	1.04	-1.67	1.17
G210	-0.22	2.05	-1.58	1.05
G215	-0.66	0.04	-1.52	1.10
G221	-0.75	0.44	-1.48	1.10
G205	-0.90	1.43	-1.36	1.10
G230	1.94	-0.67	-1.12	1.03
G202	-0.12	1.18	-1.09	1.10
G201	-1.93	-1.75	-1.02	1.00
G220	-2.27	2.08	-0.96	1.22
G219	-1.17	-0.24	-0.90	1.02
G217	-1.16	0.18	-0.89	1.08
G207	-0.06	-0.65	-0.48	1.01
G228	0.61	-0.73	-0.41	1.03
G229	0.46	-1.21	-0.30	0.99
G231	-0.97	1.95	-0.19	0.95
G213	1.37	-1.22	0.09	0.96
G211	0.15	0.71	0.28	0.99
G232	0.12	0.35	0.58	0.94
G225	0.76	2.31	0.58	0.95
G226	-0.85	1.73	0.76	0.79
G234	0.45	-0.79	0.92	0.91
G203	0.55	0.84	1.70	0.88
G233	0.55	2.24	3.12	0.76
G227	3.54	13.88	3.69	0.12
G235	1.96	5.95	4.99	0.52

TOTAL FIT-T:	Mean = -0.63 SD = 1.86 Range = -3.59 to 4.99
BETWEEN-GROUP FIT-T:	Mean = 1.51 SD = 2.70 Range = -1.75 to 13.88
DISCRIM. INDEX:	Range = -0.12 to 1.35

(iii) M1 (General Academic)

(Items ordered by total fit-t: 10 misfitting persons omitted)

ITEM NAME	ITEM DIFFIC.	BETWEEN GROUP FIT-T	TOTAL FIT-T	RASCH DISCRIM. INDEX
GA34	0.30	4.00	-5.27	1.60
GA37	0.67	4.14	-5.21	1.63
GA40	0.71	3.65	-4.96	1.59
GA24	-0.21	3.36	-4.93	1.54
GA36	0.23	3.46	-4.25	1.46
GA20	-0.38	2.75	-4.11	1.50
GA39	1.31	2.85	-3.42	1.49
GA23	-0.43	0.12	-2.42	1.25
GA33	0.77	1.96	-2.22	1.28
GA28	-0.66	0.82	-2.08	1.26
GA38	0.59	0.73	-1.97	1.25
GA32	-0.17	1.63	-1.92	1.21
GA31	-0.10	0.85	-1.61	1.26
GA07	-1.24	1.15	-1.49	1.27
GA22	-1.18	0.24	-1.40	1.21
GA14	-0.09	0.28	-1.18	1.14
GA25	0.86	-0.56	-1.02	1.11
GA35	1.25	-0.80	-0.99	1.10
GA29	0.25	-0.20	-0.83	1.12
GA21	0.73	1.61	-0.68	1.13
GA04	-0.45	-1.96	-0.67	1.06
GA18	0.78	0.41	-0.46	1.03
GA09	0.52	-0.54	-0.08	1.03
GA11	-1.79	-2.05	-0.08	0.92
GA10	-0.77	-0.93	-0.03	0.94
GA08	-1.04	2.42	0.61	0.79
GA12	-1.26	0.57	0.67	0.92
GA27	-0.17	-1.43	1.03	0.88
GA01	-1.24	2.40	1.23	0.61
GA06	-0.65	0.15	1.26	0.80
GA30	0.79	1.64	1.34	0.76
GA15	-0.32	-0.38	1.56	0.85
GA19	0.15	0.50	1.64	0.78
GA02	0.24	0.25	1.96	0.78
GA16	0.04	1.18	2.20	0.79
GA13	0.38	2.34	2.43	0.64
GA26	1.27	2.99	2.56	0.63
GA17	-0.54	2.24	3.49	0.58
GA05	0.08	5.85	6.15	0.13
GA03	0.81	10.23	9.36	0.44

TOTAL FIT-T:
Mean = -0.40
SD = 2.99
Range = -5.27 to 9.36

BETWEEN-GROUP FIT-T:
Mean = 1.45
SD = 2.27
Range = -2.05 to 10.23

DISCRIM. INDEX:
Range = -0.44 to 1.63

(iv) M1 (Life Sciences)

(Items ordered by total fit-t; 18 misfitting persons omitted)

ITEM NAME	ITEM DIFFIC.	BETWEEN GROUP FIT-T	TOTAL FIT-T	RASCH DISCRIM. INDEX
LS39	0.74	3.11	-5.07	1.65
LS40	0.95	3.53	-4.69	1.66
LS31	0.63	3.44	-4.56	1.64
LS37	0.99	2.54	-4.34	1.59
LS29	-0.06	3.14	-3.83	1.65
LS38	1.29	2.89	-3.62	1.54
LS33	0.54	2.52	-3.36	1.46
LS32	0.97	1.57	-3.32	1.46
LS27	0.80	2.70	-2.65	1.36
LS34	1.40	0.71	-2.16	1.36
LS30	0.39	1.82	-2.13	1.29
LS28	0.84	-0.11	-1.58	1.22
LS24	-1.37	0.52	-1.25	1.33
LS23	-1.47	0.58	-0.89	1.24
LS13	-1.14	-2.43	-0.32	0.98
LS25	-0.27	-0.36	-0.32	1.03
LS06	-2.51	2.21	-0.29	0.56
LS05	-2.94	-1.42	-0.19	1.13
LS26	0.53	-0.64	-0.14	1.02
LS07	-1.07	-0.13	-0.14	1.11
LS15	-0.91	0.69	-0.12	0.90
LS02	-1.87	-1.77	-0.04	0.90
LS21	-0.57	-0.59	0.00	0.90
LS35	4.18	1.32	0.18	0.47
LS18	-0.99	0.36	0.28	0.73
LS36	1.26	-0.63	0.32	0.99
LS03	-0.38	-0.78	0.34	0.84
LS12	-0.97	0.13	0.35	0.77
LS09	-1.30	-0.43	0.37	0.78
LS17	0.26	-0.82	0.51	0.92
LS20	-0.22	0.89	0.53	0.80
LS01	-1.12	0.65	0.59	0.60
LS19	-0.49	0.62	0.72	0.68
LS10	-0.09	1.01	1.67	0.70
LS11	-0.43	0.96	1.68	0.59
LS22	1.04	1.49	2.25	0.67
LS14	0.86	2.07	2.73	0.62
LS08	1.51	4.36	3.63	0.30
LS16	0.35	3.43	4.32	0.36
LS04	0.65	5.45	6.22	0.04

TOTAL FIT-T:
Mean = -0.46
SD = 2.49
Range = -5.07 to 6.22

BETWEEN-GROUP FIT-T:
Mean = 1.12
SD = 1.76
Range = -2.43 to 5.45

DISCRIM. INDEX:
Range = 0.04 to 1.66

(v) M1 (Medicine)

(Items ordered by total fit-t; 1 misfitting person omitted)

ITEM NAME	ITEM DIFFIC.	BETWEEN GROUP FIT-T	TOTAL FIT-T	RASCH DISCRIM. INDEX
ME34	0.26	1.35	-2.06	1.57
ME39	1.22	1.34	-1.76	1.37
ME23	0.49	0.48	-1.55	1.43
ME36	0.75	0.54	-1.53	1.38
ME27	0.71	1.00	-1.53	1.31
ME31	-0.52	1.06	-1.37	1.54
ME30	1.05	0.66	-0.91	1.17
ME29	-0.07	0.53	-0.87	1.28
ME03	-0.47	-0.71	-0.85	1.25
ME38	0.10	1.35	-0.82	1.29
ME33	-1.28	0.93	-0.81	1.33
ME19	-0.64	0.94	-0.79	1.23
ME09	-0.89	0.30	-0.70	1.13
ME10	-0.47	0.31	-0.67	1.15
ME28	1.15	-1.20	-0.59	1.18
ME02	-1.57	0.03	-0.56	1.16
ME17	-2.09	-0.03	-0.54	1.46
ME37	0.10	0.86	-0.33	1.08
ME40	1.05	-0.78	-0.16	0.99
ME12	-1.68	-0.24	-0.14	0.87
ME16	-0.82	-0.41	-0.05	0.96
ME01	-1.19	-0.56	-0.04	1.13
ME15	-0.12	-1.09	-0.03	0.93
ME11	-1.57	0.28	0.04	0.76
ME21	1.35	-0.59	0.07	0.92
ME18	-1.68	-0.09	0.09	0.95
ME24	1.02	-2.19	0.09	1.01
ME07	-0.82	-0.48	0.13	0.75
ME08	-0.96	1.74	0.22	0.70
ME26	1.12	0.55	0.22	0.89
ME25	0.95	0.43	0.52	0.85
ME14	-0.47	0.25	0.54	0.77
ME22	-0.70	0.35	0.56	0.71
ME32	0.92	-1.16	0.65	0.86
ME13	2.30	-0.08	0.93	0.73
ME04	-0.17	0.99	1.15	0.51
ME06	1.93	0.31	1.21	0.71
ME20	-0.52	1.67	1.40	0.60
ME05	0.34	2.46	1.69	0.38
ME35	1.83	1.39	2.59	0.44

TOTAL FIT-T:
Mean = -0.16
SD = 0.99
Range = -0.26 to 2.59

BETWEEN-GROUP FIT-T:
Mean = 0.31
SD = 0.94
Range = -2.19 to 2.46

DISCRIM. INDEX:
Range = 0.38 to 1.57

(vi) M1 (Physical Sciences)

(Items ordered by total fit-t; 1 misfitting persons omitted)

ITEM NAME	ITEM DIFFIC.	BETWEEN GROUP FIT-T	TOTAL FIT-T	RASCH DISCRIM. INDEX
PS13	1.21	0.74	-2.14	1.34
PS29	0.85	-0.13	-1.39	1.30
PS38	0.54	0.53	-1.37	1.32
PS37	0.45	1.12	-1.29	1.27
PS21	0.05	-0.69	-1.24	1.20
PS31	0.93	0.52	-1.05	1.22
PS26	-0.00	0.02	-0.93	1.22
PS34	-0.78	0.84	-0.92	1.33
PS16	-1.58	-0.95	-0.89	1.13
PS02	0.35	0.07	-0.82	1.13
PS09	-2.26	-1.12	-0.80	1.23
PS33	0.31	-1.56	-0.69	1.16
PS10	-0.11	-0.28	-0.64	1.09
PS03	-0.78	-1.19	-0.63	1.15
PS11	-1.72	-0.24	-0.56	0.97
PS19	-1.34	-0.77	-0.52	1.13
PS36	0.21	-0.17	-0.46	1.15
PS05	-0.29	-0.89	-0.45	1.09
PS23	-0.63	-0.09	-0.42	1.12
PS08	-1.23	0.33	-0.33	0.79
PS30	-0.86	0.26	-0.30	1.21
PS27	-0.94	0.01	-0.27	1.17
PS04	2.28	-0.54	-0.19	0.98
PS28	-0.00	0.35	-0.11	1.05
PS01	-1.72	0.87	-0.09	0.75
PS20	-0.94	-1.51	0.03	0.98
PS07	1.65	-1.63	0.10	0.96
PS22	-1.23	1.45	0.12	0.92
PS17	-0.55	0.44	0.24	0.84
PS14	0.31	0.70	0.40	0.94
PS24	-0.35	1.02	0.72	0.77
PS39	1.05	0.12	0.74	0.89
PS25	-0.29	0.39	0.86	0.91
PS12	0.59	0.35	0.86	0.78
PS18	-0.29	-0.43	0.98	0.82
PS35	1.29	-1.23	1.09	0.85
PS15	0.31	0.72	1.25	0.73
PS32	0.67	1.83	1.35	0.70
PS40	3.81	3.44	1.71	0.32
PS06	1.05	1.04	2.32	0.49

TOTAL FIT-T:
Mean = -0.14
SD = 0.95
Range = -2.14 to 2.32

BETWEEN-GROUP FIT-T:
Mean = 0.09
SD = 1.01
Range = -1.63 to 3.44

DISCRIM. INDEX:
Range = 0.32 to 1.34

(vii) M1 (Social Studies)

(Items ordered by total fit-t; 7 misfitting persons omitted)

ITEM NAME	ITEM DIFFIC.	BETWEEN GROUP FIT-T	TOTAL FIT-T	RASCH DISCRIM. INDEX
SS33	-0.13	1.71	-3.46	1.55
SS32	0.40	1.29	-2.71	1.51
SS11	0.03	1.69	-2.63	1.51
SS28	-0.32	1.20	-2.33	1.50
SS30	0.62	1.19	-2.17	1.43
SS31	0.10	2.01	-2.09	1.43
SS37	-0.26	0.92	-1.96	1.40
SS38	0.88	-0.31	-1.41	1.27
SS12	0.78	-0.14	-1.38	1.22
SS27	-0.73	0.58	-1.34	1.28
SS01	0.37	0.11	-1.11	1.16
SS05	1.19	0.69	-1.10	1.24
SS18	-1.90	0.72	-0.97	1.47
SS23	0.17	0.43	-0.90	1.11
SS35	0.07	-0.65	-0.88	1.15
SS13	-0.59	1.51	-0.85	1.26
SS29	0.82	-0.44	-0.69	1.12
SS09	-1.30	-0.29	-0.56	1.18
SS07	-0.51	-0.88	-0.53	1.05
SS03	-0.11	0.11	-0.41	1.14
SS17	-2.59	0.73	-0.38	0.84
SS08	0.84	-0.20	-0.30	1.10
SS34	0.74	-0.86	-0.21	0.99
SS02	-0.26	-0.96	-0.19	1.14
SS16	-2.36	-1.24	-0.11	0.91
SS39	0.49	-1.86	-0.04	0.98
SS15	-1.30	-0.76	0.31	0.91
SS21	0.46	-1.96	0.55	0.87
SS26	0.05	0.62	0.74	0.83
SS10	-0.15	0.24	0.76	0.89
SS20	-0.47	0.73	0.92	0.73
SS40	1.43	1.69	1.09	0.65
SS06	-0.22	1.67	1.11	0.76
SS25	-0.08	-0.91	1.18	0.75
SS36	1.25	1.20	1.50	0.53
SS04	-0.59	1.15	2.01	0.54
SS19	2.23	4.30	2.05	0.06
SS24	0.28	1.68	2.65	0.49
SS22	0.71	2.05	2.65	0.42
SS14	-0.02	2.19	3.18	0.35

TOTAL FIT-T:
Mean = -0.25
SD = 1.57
Range = -3.46 to 3.18

BETWEEN-GROUP FIT-T:
Mean = 0.52
SD = 1.26
Range = -1.96 to 4.30

DISCRIM. INDEX:
Range = 0.06 to 1.55

(viii) M1 (Technology)

(Items ordered by total fit-t; 2 misfitting persons omitted)

ITEM NAME	ITEM DIFFIC.	BETWEEN GROUP FIT-T	TOTAL FIT-T	RASCH DISCRIM. INDEX
TN36	0.23	1.57	-2.00	1.46
TN34	-0.64	2.75	-1.85	1.53
TN03	-1.63	0.28	-1.81	1.38
TN32	-1.44	0.92	-1.80	1.48
TN15	-1.20	0.68	-1.70	1.46
TN08	-1.73	0.21	-1.61	1.17
TN31	-0.14	2.57	-1.57	1.32
TN23	0.67	0.59	-1.49	1.19
TN04	-1.36	0.08	-1.41	1.37
TN07	-1.44	0.41	-1.34	1.35
TN16	-0.49	1.27	-1.20	1.34
TN35	1.24	0.42	-1.10	1.15
TN25	-0.81	-0.16	-1.06	1.24
TN10	0.40	1.73	-1.01	1.27
TN33	0.92	-0.30	-0.98	1.08
TN17	0.55	0.55	-0.91	1.22
TN40	0.95	-0.27	-0.69	1.20
TN14	-0.54	-1.48	-0.67	1.14
TN02	-0.22	-1.66	-0.64	0.99
TN20	0.02	-0.19	-0.61	1.12
TN26	-0.14	-0.68	-0.58	1.06
TN29	-0.18	-1.23	-0.52	1.04
TN18	0.30	-0.54	-0.42	1.16
TN09	-0.75	-0.70	-0.30	0.86
TN12	0.90	-0.36	-0.22	1.01
TN05	0.14	-0.32	-0.09	0.99
TN01	0.75	1.38	-0.04	0.52
TN19	0.16	-1.00	0.20	1.08
TN30	0.16	1.22	0.30	0.75
TN11	0.79	-1.51	0.32	0.90
TN37	0.13	1.05	0.37	1.22
TN21	1.22	0.75	0.49	0.91
TN06	-1.96	1.24	0.59	0.74
TN39	1.87	-1.36	0.93	0.92
TN22	0.90	1.80	1.91	0.71
TN28	0.61	1.35	2.01	0.80
TN24	0.49	1.86	2.09	0.39
TN38	0.98	1.94	2.20	0.53
TN27	0.52	3.10	3.32	0.40
TN13	1.56	4.16	6.01	0.02

TOTAL FIT-T:
Mean = -0.17
SD = 1.61
Range = -2.00 to 6.01

BETWEEN-GROUP FIT-T:
Mean = 0.55
SD = 1.35
Range = -1.66 to 4.16

DISCRIM. INDEX:
Range = -0.02 to 1.53

I.5 ELTS Subtests: Ability/Difficulty Scales

(i) G1 (Reading)

PERSON STATS	COUNT	RAW SCORE	MEASURE MIDPOINT	ITEM COUNTS	ITEM NAMES			
	40	39	3.90					
			3.70					
			3.50					
+2SD	58	38	3.30					
			3.10					
	61	37	2.90					
			2.70					
	84	36	2.50					
+1SD	75	35	2.30					
			2.10					
	74	34	1.90					
	56	33	1.70	2	G111	G138		
	129	31	1.50	2	G125	G126		
MEAN	86	30	1.30	2	G112	G135		
	62	29	1.10	1	G137			
	163	27	0.90	1	G132			
	72	26	0.70	4	G109	G110	G117	G140
	141	24	0.50	2	G131	G139		
	107	22	0.30	3	G118	G124	G128	
-1SD	36	21	0.10	3	G130	G134	G136	
	66	19	-0.10	2	G114	G115		
	55	17	-0.30	4	G116	G121	G123	G129
	32	16	-0.50	2	G103	G133		
	22	14	-0.70	3	G105	G107	G122	
-2SD	14	12	-0.90	2	G108	G127		
	3	11	-1.10	2	G106	G113		
	4	10	-1.30	2	G104	G120		
	5	8	-1.50	1	G119			
	3	7	-1.70	1	G102			
-3SD		6	-1.90	1	G101			
			-2.10					

40 ITEMS CALIBRATED ON 1448 PERSONS

(ii) G2 (Listening)

PERSON STATS	COUNT	RAW SCORE	MEASURE MIDPOINT	ITEM COUNTS	ITEM NAMES				
	11	34	4.10						
			3.90						
			3.70						
+3SD			3.50	1	Q227				
	37	33	3.30						
			3.10						
			2.90						
+2SD	41	32	2.70						
	41	31	2.50						
			2.30						
	83	30	2.10						
+1SD	77	29	1.90	2	Q230	Q235			
	78	28	1.70						
	107	27	1.50						
	118	26	1.30	1	Q213				
	104	25	1.10						
MEAN	103	24	0.90						
	99	23	0.70	5	Q204	Q214	Q216	Q225	Q228
	191	21	0.50	4	Q203	Q229	Q233	Q234	
	63	20	0.30	1	Q209				
-1SD	131	18	0.10	3	Q211	Q222	Q232		
	49	17	-0.10	4	Q202	Q206	Q207	Q208	
	66	15	-0.30	2	Q210	Q223			
	18	14	-0.50						
-2SD	18	13	-0.70	3	Q212	Q215	Q221		
	30	11	-0.90	4	Q205	Q224	Q226	Q231	
	5	10	-1.10	2	Q217	Q219			
	5	9	-1.30						
-3SD	3	8	-1.50						
	1	7	-1.70	1	Q218				
		6	-1.90	1	Q201				
	1	5	-2.10						
	1	4	-2.30	1	Q220				
-4SD			-2.50						

35 ITEMS CALIBRATED ON 1481 PERSONS

(iii) M1 (General Academic)

PERSON STATS	COUNT	RAW SCORE	MEASURE MIDPOINT	ITEM COUNTS	ITEM NAMES					
	2	39	3.90							
			3.70							
			3.50							
+3SD			3.30							
	3	38	3.10							
			2.90							
	6	37	2.70							
			2.50							
+2SD	12	36	2.30							
	6	35	2.10							
	8	34	1.90							
	8	33	1.70							
	12	32	1.50							
+1SD	17	30	1.30	3	CA26	CA35	CA39			
	9	29	1.10							
	28	27	0.90	2	CA03	CA25				
	13	26	0.70	6	CA18	CA21	CA30	CA33	CA37	CA40
	19	24	0.50	2	CA09	CA38				
	30	22	0.30	5	CA02	CA13	CA29	CA34	CA36	
MEAN	38	20	0.10	3	CA05	CA16	CA19			
	16	19	-0.10	4	CA14	CA27	CA31	CA32		
	32	17	-0.30	3	CA15	CA20	CA24			
	38	15	-0.50	3	CA04	CA17	CA23			
	32	13	-0.70	3	CA06	CA10	CA28			
-1SD	19	12	-0.90							
	14	11	-1.10	2	CA08	CA22				
	15	9	-1.30	3	CA01	CA07	CA12			
	3	8	-1.50							
	5	7	-1.70	1	CA11					
-2SD	3	6	-1.90							
	4	5	-2.10							

40 ITEMS CALIBRATED ON 392 PERSONS

(iv) M1 (Life Sciences)

PERSON STATS	COUNT	RAW SCORE	MEASURE MIDPOINT	ITEM COUNTS	ITEM NAMES					
			4.10	1	LS35					
+4SD			3.90							
			3.70							
		38	3.50							
			3.30							
+3SD	3	37	3.10							
			2.90							
	10	36	2.70							
			2.50							
+2SD	7	35	2.30							
	7	34	2.10							
	14	33	1.90							
	14	32	1.70							
+1SD	13	31	1.50	1	LS08					
	17	30	1.30	3	LS34	LS36	LS38			
	44	28	1.10	1	LS22					
	24	27	0.90	6	LS14	LS27	LS28	LS32	LS37	LS40
MEAN	32	25	0.70	3	LS04	LS31	LS39			
	16	24	0.50	2	LS26	LS33				
	38	22	0.30	3	LS16	LS17	LS30			
	16	21	0.10							
-1SD	47	19	-0.10	2	LS10	LS29				
	20	17	-0.30	3	LS03	LS20	LS25			
	11	16	-0.50	3	LS11	LS19	LS21			
	14	14	-0.70							
-2SD	2	13	-0.90	3	LS12	LS15	LS18			
		12	-1.10	3	LS01	LS07	LS13			
	4	10	-1.30	2	LS09	LS24				
	1	9	-1.50	1	LS23					
-3SD	1	8	-1.70							
		7	-1.90	1	LS02					
		6	-2.10							
		5	-2.30							
-4SD			-2.50	1	LS06					
		4	-2.70							
			-2.90	1	LS05					
		3	-3.10							

40 ITEMS CALIBRATED ON 355 PERSONS

(v) M1 (Medicine)

PERSON STATS	COUNT	RAW SCORE	MEASURE MIDPOINT	ITEM COUNTS	ITEM NAMES				
	3	38	3.30						
			3.10						
+2SD			2.90						
	6	37	2.70						
			2.50						
	6	36	2.30	1	ME13				
	6	35	2.10						
+1SD	15	34	1.90	2	ME06	ME35			
	12	33	1.70						
	8	32	1.50						
	9	31	1.30	2	ME21	ME39			
MEAN	9	30	1.10	5	ME24	ME26	ME28	ME30	ME40
	15	28	0.90	2	ME25	ME32			
	8	27	0.70	2	ME27	ME36			
	12	25	0.50	1	ME23				
	4	24	0.30	2	ME05	ME34			
-1SD	11	22	0.10	2	ME37	ME38			
	4	21	-0.10	3	ME04	ME15	ME29		
	6	19	-0.30						
	1	17	-0.50	5	ME03	ME10	ME14	ME20	ME31
-2SD	2	16	-0.70	2	ME19	ME22			
	1	14	-0.90	4	ME07	ME08	ME09	ME16	
		13	-1.10	1	ME01				
	1	11	-1.30	1	ME33				
-3SD		10	-1.50	2	ME02	ME11			
	2	9	-1.70	2	ME12	ME18			
	1	8	-1.90						
		7	-2.10	1	ME17				
		6	-2.30						
-4SD		5							

40 ITEMS CALIBRATED ON 142 PERSONS

(vi) M1 (Physical Sciences)

PERSON STATS	COUNT	RAW SCORE	MEASURE MIDPOINT	ITEM COUNTS	ITEM NAMES				
	3	39	4.30						
			4.10						
+2SD			3.90	1	PS40				
			3.70						
	8	38	3.50						
			3.30						
			3.10						
	12	37	2.90						
+1SD	9	36	2.70						
			2.50						
	9	35	2.30	1	PS04				
	11	34	2.10						
	8	33	1.90						
	11	32	1.70	1	PS07				
MEAN	8	31	1.50						
	4	30	1.30	2	PS13	PS35			
	4	29	1.10	2	PS06	PS39			
	12	27	0.90	2	PS29	PS31			
	2	26	0.70	1	PS32				
-1SD	6	24	0.50	3	PS12	PS37	PS38		
	6	22	0.30	5	PS02	PS14	PS15	PS33	PS36
	6	21	0.10	1	PS21				
	5	19	-0.10	3	PS10	PS26	PS28		
		18	-0.30	4	PS05	PS18	PS24	PS25	
	4	16	-0.50	1	PS17				
-2SD	1	14	-0.70	3	PS03	PS23	PS34		
		13	-0.90	3	PS20	PS27	PS30		
		11	-1.10						
	1	10	-1.30	3	PS08	PS19	PS22		
		9	-1.50	1	PS16				
-3SD	2	8	-1.70	2	PS01	PS11			
		7	-1.90						
		6	-2.10						
		5	-2.30	1	PS09				

40 ITEMS CALIBRATED ON 132 PERSONS

(vii) M1 (Social Studies)

PERSON STATS	COUNT	RAW SCORE	MEASURE MIDPOINT	ITEM COUNTS	ITEM NAMES				
	1	39	3.90						
			3.70						
+4SD			3.50						
	1	38	3.30						
			3.10						
			2.90						
+3SD	1	37	2.70						
	1	36	2.50						
			2.30	1	SS19				
	3	35	2.10						
+2SD	7	34	1.90						
	4	33	1.70						
	11	31	1.50	1	SS40				
	8	30	1.30	1	SS36				
+1SD	9	29	1.10	1	SS05				
	15	27	0.90	3	SS08	SS29	SS38		
	30	25	0.70	4	SS12	SS22	SS30	SS34	
	17	24	0.50	3	SS21	SS32	SS39		
MEAN	29	22	0.30	2	SS01	SS24			
	31	20	0.10	5	SS11	SS23	SS26	SS31	SS35
	15	19	-0.10	5	SS03	SS10	SS14	SS25	SS33
	18	17	-0.30	4	SS02	SS06	SS28	SS37	
-1SD	26	15	-0.50	4	SS04	SS07	SS13	SS20	
	5	14	-0.70	1	SS27				
	10	12	-0.90						
	5	11	-1.10						
-2SD	5	9	-1.30	2	SS09	SS15			
	2	8	-1.50						
		7	-1.70						
	3	6	-1.90	1	SS18				
-3SD			-2.10						
		5	-2.30	1	SS16				
		4	-2.50	1	SS17				
			-2.70						

40 ITEMS CALIBRATED ON 257 PERSONS

(viii) M1 (Technology)

PERSON STATS	COUNT	RAW SCORE	MEASURE MIDPOINT	ITEM COUNTS	ITEM NAMES				
	4	39	3.90						
			3.70						
			3.50						
+2SD	4	38	3.30						
			3.10						
			2.90						
	8	37	2.70						
	12	36	2.50						
+1SD			2.30						
	15	35	2.10						
	13	34	1.90	1	TN39				
	20	33	1.70						
	25	31	1.50	1	TN13				
MEAN	13	30	1.30	2	TN21	TN35			
	14	29	1.10						
	14	27	0.90	5	TN12	TN22	TN33	TN38	TN40
	14	25	0.70	3	TN11	TN23	TN28		
-1SD	4	24	0.50	3	TN17	TN24	TN27		
	4	22	0.30	3	TN10	TN18	TN36		
	3	20	0.10	4	TN19	TN20	TN30	TN37	
	3	19	-0.10	4	TN05	TN26	TN29	TN31	
	3	17	-0.30	1	TN02				
-2SD		15	-0.50	2	TN14	TN16			
	1	14	-0.70	3	TN01	TN09	TN34		
		12	-0.90	1	TN25				
		11	-1.10						
	1	10	-1.30	2	TN04	TN15			
-3SD	1	8	-1.50	2	TN07	TN32			
	2	7	-1.70	2	TN03	TN08			
	4	6	-1.90	1	TN06				
			-2.10						

40 ITEMS CALIBRATED ON 182 PERSONS

APPENDIX J
ELTS DATA SUBSETS & COMBINED SUBTESTS: RASCH DIFFICULTIES

J.1 Difficulty estimates from Combined Calibration of G1 & G2

ITEM NAME	ITEM DIFFIC.	STANDARD ERROR
G101	-1.93	0.11
G102	-1.72	0.10
G103	-0.59	0.07
G104	-1.31	0.09
G105	-0.79	0.07
G106	-1.08	0.08
G107	-0.74	0.07
G108	-1.07	0.08
G109	0.66	0.06
G110	0.53	0.06
G111	1.52	0.06
G112	1.11	0.06
G113	-1.16	0.08
G114	-0.29	0.07
G115	-0.21	0.06
G116	-0.31	0.07
G117	0.65	0.06
G118	0.16	0.06
G119	-1.47	0.09
G120	-1.25	0.09
G121	-0.39	0.07
G122	-0.76	0.07
G123	-0.36	0.07
G124	0.27	0.06
G125	1.43	0.06
G126	1.35	0.06
G127	-1.04	0.08
G128	0.11	0.06
G129	-0.30	0.07
G130	0.08	0.06
G131	0.49	0.06
G132	0.70	0.06
G133	-0.60	0.07
G134	-0.08	0.06
G135	1.09	0.06
G136	0.02	0.06
G137	0.88	0.06
G138	1.56	0.06
G139	0.37	0.06
G140	0.59	0.06

ITEM NAME	ITEM DIFFIC.	STANDARD ERROR
G201	-1.82	0.11
G202	-0.03	0.06
G203	0.66	0.06
G204	0.77	0.06
G205	-0.79	0.07
G206	-0.09	0.06
G207	0.07	0.06
G208	-0.04	0.06
G209	0.37	0.06
G210	-0.10	0.06
G211	0.25	0.06
G212	-0.55	0.07
G213	1.49	0.06
G214	0.81	0.06
G215	-0.57	0.07
G216	0.87	0.06
G217	-1.06	0.08
G218	-1.55	0.10
G219	-1.03	0.08
G220	-2.15	0.12
G221	-0.65	0.07
G222	0.12	0.06
G223	-0.27	0.07
G224	-0.77	0.07
G225	0.88	0.06
G226	-0.75	0.07
G227	3.77	0.10
G228	0.73	0.06
G229	0.56	0.06
G230	2.08	0.06
G231	-0.90	0.08
G232	0.24	0.06
G233	0.65	0.06
G234	0.55	0.06
G235	2.10	0.06

Items calibrated on 1,465 persons

J.2 Difficulty Estimates from Combined Calibration of G1 + G2 + M1(GA)

ITEM NAME	ITEM DIFFIC.	STANDARD ERROR
G101	-2.11	0.20
G102	-2.20	0.21
G103	-1.01	0.14
G104	-1.63	0.17
G105	-1.22	0.15
G106	-1.12	0.15
G107	-1.49	0.16
G108	-1.54	0.17
G109	0.31	0.11
G110	0.31	0.11
G111	1.19	0.11
G112	0.72	0.11
G113	-1.49	0.16
G114	-0.72	0.13
G115	-1.05	0.14
G116	-1.01	0.14
G117	0.20	0.11
G118	-0.34	0.12
G119	-1.75	0.18
G120	-1.44	0.16
G121	-1.03	0.14
G122	-1.14	0.15
G123	-0.63	0.13
G124	-0.08	0.12
G125	1.05	0.11
G126	1.14	0.11
G127	-1.27	0.15
G128	-0.19	0.12
G129	-0.59	0.13
G130	-0.31	0.12
G131	0.04	0.12
G132	0.21	0.11
G133	-0.95	0.14
G134	-0.51	0.13
G135	0.76	0.11
G136	-0.46	0.12
G137	0.35	0.11
G138	1.09	0.11
G139	-0.18	0.12
G140	0.03	0.12
G201	-1.82	0.18
G202	-0.18	0.12
G203	0.47	0.11
G204	0.67	0.11
G205	-0.86	0.14
G206	-0.21	0.12
G207	0.12	0.12
G208	0.09	0.12
G209	0.08	0.12
G210	-0.19	0.12
G211	0.11	0.12
G212	-0.62	0.13
G213	1.24	0.11
G214	0.69	0.11
G215	-0.81	0.13
G216	0.48	0.11
G217	-1.36	0.16
G218	-2.20	0.21
G219	-1.44	0.16
G220	-2.75	0.27
G221	-1.34	0.16
G222	-0.59	0.13
G223	-1.12	0.15
G224	-1.34	0.16
G225	0.76	0.11

ITEM NAME	ITEM DIFFIC.	STANDARD ERROR
G226	-1.18	0.15
G227	3.64	0.19
G228	0.34	0.11
G229	0.17	0.11
G230	1.75	0.12
G231	-1.03	0.14
G232	-0.16	0.12
G233	0.45	0.11
G234	0.31	0.11
G235	1.73	0.12
GA01	-0.62	0.13
GA02	0.89	0.11
GA03	1.48	0.12
GA04	0.20	0.11
GA05	0.75	0.11
GA06	-0.03	0.12
GA07	-0.54	0.13
GA08	-0.40	0.12
GA09	1.21	0.11
GA10	-0.09	0.12
GA11	-1.07	0.14
GA12	-0.57	0.13
GA13	1.03	0.11
GA14	0.58	0.11
GA15	0.38	0.11
GA16	0.70	0.11
GA17	0.11	0.12
GA18	1.44	0.12
GA19	0.78	0.11
GA20	0.24	0.11
GA21	1.37	0.12
GA22	-0.47	0.13
GA23	0.21	0.11
GA24	0.44	0.11
GA25	1.49	0.12
GA26	1.94	0.12
GA27	0.45	0.11
GA28	0.01	0.12
GA29	0.88	0.11
GA30	1.47	0.12
GA31	0.63	0.11
GA32	0.49	0.11
GA33	1.41	0.12
GA34	0.95	0.11
GA35	1.83	0.12
GA36	0.81	0.11
GA37	1.31	0.11
GA38	1.22	0.11
GA39	1.89	0.12
GA40	1.32	0.11

Items calibrated on 390 persons

J.3 Difficulty Estimates from Combined Calibration of G1 + G2 + M1(LS)

ITEM NAME	ITEM DIFFIC.	STANDARD ERROR
G101	-2.38	0.25
G102	-1.76	0.19
G103	-0.89	0.14
G104	-1.34	0.17
G105	-1.11	0.15
G106	-1.49	0.18
G107	-0.63	0.14
G108	-1.31	0.17
G109	0.53	0.11
G110	0.55	0.11
G111	1.29	0.12
G112	1.04	0.11
G113	-1.31	0.17
G114	-0.28	0.13
G115	-0.16	0.12
G116	-0.33	0.13
G117	0.66	0.11
G118	-0.15	0.12
G119	-1.58	0.18
G120	-1.46	0.17
G121	-0.31	0.13
G122	-0.73	0.14
G123	-0.51	0.13
G124	0.22	0.12
G125	1.25	0.11
G126	1.31	0.12
G127	-1.34	0.17
G128	-0.41	0.13
G129	-0.36	0.13
G130	0.04	0.12
G131	0.57	0.11
G132	0.84	0.11
G133	-0.41	0.13
G134	-0.00	0.12
G135	1.18	0.11
G136	0.26	0.12
G137	0.95	0.11
G138	1.61	0.12
G139	0.57	0.11
G140	0.77	0.11
G201	-2.38	0.25
G202	-0.48	0.13
G203	0.42	0.11
G204	0.76	0.11
G205	-0.71	0.14
G206	-0.09	0.12
G207	-0.16	0.12
G208	-0.27	0.12
G209	0.01	0.12
G210	0.06	0.12
G211	0.04	0.12
G212	-0.87	0.14
G213	1.30	0.12
G214	0.62	0.11
G215	-0.77	0.14
G216	0.87	0.11
G217	-1.06	0.15
G218	-1.61	0.18
G219	-0.85	0.14
G220	-2.32	0.24
G221	-0.77	0.14
G222	0.03	0.12
G223	-0.21	0.12
G224	-0.73	0.14
G225	0.62	0.11

ITEM NAME	ITEM DIFFIC.	STANDARD ERROR
G226	-0.63	0.14
G227	2.89	0.16
G228	0.70	0.11
G229	0.57	0.11
G230	1.82	0.12
G231	-1.13	0.16
G232	0.27	0.12
G233	0.62	0.11
G234	0.48	0.11
G235	1.80	0.12
LS01	-0.81	0.14
LS02	-1.61	0.18
LS03	-0.22	0.12
LS04	0.82	0.11
LS05	-2.51	0.27
LS06	-2.15	0.23
LS07	-0.77	0.14
LS08	1.63	0.12
LS09	-0.97	0.15
LS10	0.06	0.12
LS11	-0.27	0.12
LS12	-0.75	0.14
LS13	-0.95	0.15
LS14	1.01	0.11
LS15	-0.71	0.14
LS16	0.55	0.11
LS17	0.36	0.11
LS18	-0.79	0.14
LS19	-0.30	0.13
LS20	-0.02	0.12
LS21	-0.35	0.13
LS22	1.18	0.11
LS23	-1.26	0.16
LS24	-1.18	0.16
LS25	-0.06	0.12
LS26	0.71	0.11
LS27	0.93	0.11
LS28	1.00	0.11
LS29	0.12	0.12
LS30	0.55	0.11
LS31	0.80	0.11
LS32	1.14	0.11
LS33	0.75	0.11
LS34	1.50	0.12
LS35	4.38	0.29
LS36	1.42	0.12
LS37	1.10	0.11
LS38	1.42	0.12
LS39	0.90	0.11
LS40	1.08	0.11

Items calibrated on 356 persons

J.4 Difficulty Estimates from Combined Calibration of G1 + G2 + M1(ME)

ITEM NAME	ITEM DIFFIC.	STANDARD ERROR
G101	-2.81	0.59
G102	-2.27	0.46
G103	-0.20	0.23
G104	-1.90	0.40
G105	-0.47	0.24
G106	-0.79	0.27
G107	-0.53	0.25
G108	-0.65	0.26
G109	0.68	0.19
G110	-0.10	0.22
G111	2.19	0.19
G112	1.37	0.18
G113	-1.75	0.38
G114	-0.41	0.24
G115	0.04	0.21
G116	-0.47	0.24
G117	0.61	0.19
G118	0.21	0.21
G119	-1.39	0.33
G120	-0.86	0.27
G121	-0.30	0.23
G122	-1.39	0.33
G123	-0.36	0.23
G124	0.34	0.20
G125	1.60	0.18
G126	0.75	0.19
G127	-1.19	0.30
G128	0.30	0.20
G129	-0.41	0.24
G130	-0.25	0.23
G131	0.30	0.20
G132	0.89	0.19
G133	-0.53	0.25
G134	-0.41	0.24
G135	0.93	0.19
G136	-0.20	0.23
G137	0.64	0.19
G138	1.23	0.18
G139	0.13	0.21
G140	-0.15	0.22
G201	-2.81	0.59
G202	-0.15	0.22
G203	0.96	0.19
G204	0.49	0.20
G205	-1.02	0.29
G206	-0.41	0.24
G207	-0.05	0.22
G208	0.04	0.21
G209	0.45	0.20
G210	-0.47	0.24
G211	0.42	0.20
G212	-0.86	0.27
G213	1.53	0.18
G214	0.53	0.20
G215	-0.47	0.24
G216	0.93	0.19
G217	-1.29	0.31
G218	-1.29	0.31
G219	-1.39	0.33
G220	-2.07	0.43
G221	-0.53	0.25
G222	0.57	0.19
G223	-0.05	0.22
G224	-0.79	0.27
G225	0.82	0.19

ITEM NAME	ITEM DIFFIC.	STANDARD ERROR
G226	-0.94	0.28
G227	4.54	0.37
G228	0.89	0.19
G229	0.53	0.20
G230	2.22	0.19
G231	-1.19	0.30
G232	-0.25	0.23
G233	0.42	0.20
G234	0.53	0.20
G235	2.15	0.19
ME01	-1.02	0.29
ME02	-1.39	0.33
ME03	-0.36	0.23
ME04	-0.01	0.22
ME05	0.45	0.20
ME06	2.04	0.19
ME07	-0.65	0.26
ME08	-0.79	0.27
ME09	-0.72	0.26
ME10	-0.30	0.23
ME11	-1.39	0.33
ME12	-1.50	0.34
ME13	2.45	0.20
ME14	-0.30	0.23
ME15	0.04	0.21
ME16	-0.72	0.26
ME17	-1.90	0.40
ME18	-1.50	0.34
ME19	-0.53	0.25
ME20	-0.41	0.24
ME21	1.47	0.18
ME22	-0.53	0.25
ME23	0.61	0.19
ME24	1.17	0.18
ME25	1.06	0.19
ME26	1.23	0.18
ME27	0.82	0.19
ME28	1.27	0.18
ME29	0.04	0.21
ME30	1.20	0.18
ME31	-0.41	0.24
ME32	1.03	0.19
ME33	-1.10	0.29
ME34	0.42	0.20
ME35	1.94	0.19
ME36	0.86	0.19
ME37	0.25	0.20
ME38	0.25	0.20
ME39	1.37	0.18
ME40	1.20	0.18

Items calibrated on 141 persons

J.5 Difficulty Estimates from Combined Calibration of G1 + G2 + M1(PS)

ITEM NAME	ITEM DIFFIC.	STANDARD ERROR
G101	-1.17	0.30
G102	-1.70	0.36
G103	-0.30	0.23
G104	-1.46	0.33
G105	-0.47	0.24
G106	-1.26	0.31
G107	-0.41	0.24
G108	-1.09	0.29
G109	0.84	0.20
G110	0.32	0.21
G111	1.57	0.20
G112	0.92	0.20
G113	-0.65	0.25
G114	-0.19	0.23
G115	0.05	0.22
G116	-0.41	0.24
G117	0.69	0.20
G118	0.40	0.21
G119	-1.46	0.33
G120	-1.09	0.29
G121	-0.59	0.25
G122	-1.09	0.29
G123	-0.30	0.23
G124	0.14	0.21
G125	1.72	0.20
G126	1.41	0.20
G127	-0.47	0.24
G128	0.40	0.21
G129	-0.14	0.23
G130	0.19	0.21
G131	0.69	0.20
G132	1.04	0.20
G133	-0.35	0.24
G134	0.23	0.21
G135	1.38	0.20
G136	0.36	0.21
G137	1.04	0.20
G138	1.84	0.20
G139	0.69	0.20
G140	1.07	0.20
G201	-1.36	0.32
G202	0.28	0.21
G203	0.57	0.20
G204	0.88	0.20
G205	-1.17	0.30
G206	0.00	0.22
G207	-0.19	0.23
G208	-0.47	0.24
G209	0.49	0.20
G210	-0.30	0.23
G211	0.40	0.21
G212	-0.19	0.23
G213	1.45	0.20
G214	1.26	0.19
G215	-0.59	0.25
G216	0.65	0.20
G217	-0.65	0.25
G218	-1.09	0.29
G219	-0.53	0.25
G220	-1.57	0.34
G221	-0.59	0.25
G222	0.61	0.20
G223	0.00	0.22
G224	-0.25	0.23
G225	0.88	0.20

ITEM NAME	ITEM DIFFIC.	STANDARD ERROR
G226	-0.53	0.25
G227	4.00	0.32
G228	0.81	0.20
G229	0.92	0.20
G230	2.00	0.20
G231	-0.93	0.28
G232	0.40	0.21
G233	0.96	0.20
G234	0.05	0.22
G235	2.38	0.21
PS01	-1.83	0.38
PS02	0.10	0.22
PS03	-1.17	0.30
PS04	1.92	0.20
PS05	-0.59	0.25
PS06	0.77	0.20
PS07	1.30	0.19
PS08	-1.36	0.32
PS09	-2.59	0.52
PS10	-0.41	0.24
PS11	-1.98	0.40
PS12	0.28	0.21
PS13	0.88	0.20
PS14	0.00	0.22
PS15	0.00	0.22
PS16	-1.70	0.36
PS17	-0.78	0.26
PS18	-0.41	0.24
PS19	-1.57	0.34
PS20	-1.17	0.30
PS21	-0.25	0.23
PS22	-1.36	0.32
PS23	-0.85	0.27
PS24	-0.65	0.25
PS25	-0.53	0.25
PS26	-0.25	0.23
PS27	-1.26	0.31
PS28	-0.30	0.23
PS29	0.57	0.20
PS30	-1.17	0.30
PS31	0.69	0.20
PS32	0.36	0.21
PS33	0.00	0.22
PS34	-1.00	0.28
PS35	1.00	0.20
PS36	-0.14	0.23
PS37	0.19	0.21
PS38	0.23	0.21
PS39	0.69	0.20
PS40	3.35	0.26

Items calibrated on 130 persons

J.6 Difficulty Estimates from Combined Calibration of G1 + G2 + M1(SS)

ITEM NAME	ITEM DIFFIC.	STANDARD ERROR
G101	-1.90	0.24
G102	-1.96	0.25
G103	-0.70	0.16
G104	-1.41	0.20
G105	-0.81	0.17
G106	-1.12	0.18
G107	-0.96	0.18
G108	-1.59	0.22
G109	0.38	0.14
G110	0.27	0.14
G111	1.03	0.13
G112	0.94	0.13
G113	-1.59	0.22
G114	-0.60	0.16
G115	-0.48	0.15
G116	-0.48	0.15
G117	0.40	0.14
G118	-0.06	0.14
G119	-1.84	0.24
G120	-2.02	0.26
G121	-0.50	0.16
G122	-0.96	0.18
G123	-0.81	0.17
G124	0.08	0.14
G125	0.98	0.13
G126	1.21	0.14
G127	-1.37	0.20
G128	-0.06	0.14
G129	-0.63	0.16
G130	-0.04	0.14
G131	-0.04	0.14
G132	0.06	0.14
G133	-1.09	0.18
G134	-0.36	0.15
G135	0.73	0.13
G136	-0.29	0.15
G137	0.60	0.13
G138	1.25	0.14
G139	0.06	0.14
G140	0.14	0.14
G201	-2.71	0.34
G202	-0.36	0.15
G203	0.51	0.14
G204	0.53	0.14
G205	-1.41	0.20
G206	-0.43	0.15
G207	-0.02	0.14
G208	-0.16	0.15
G209	0.29	0.14
G210	-0.45	0.15
G211	-0.12	0.14
G212	-0.99	0.18
G213	1.32	0.14
G214	0.64	0.13
G215	-0.73	0.16
G216	0.46	0.14
G217	-1.33	0.20
G218	-1.84	0.24
G219	-1.73	0.23
G220	-2.24	0.28
G221	-0.81	0.17
G222	-0.21	0.15
G223	-0.63	0.16
G224	-0.93	0.17
G225	0.34	0.14

ITEM NAME	ITEM DIFFIC.	STANDARD ERROR
G226	-1.19	0.19
G227	3.55	0.24
G228	0.42	0.14
G229	0.19	0.14
G230	1.67	0.14
G231	-0.96	0.18
G232	0.19	0.14
G233	0.12	0.14
G234	0.25	0.14
G235	1.93	0.15
SS01	0.92	0.13
SS02	0.27	0.14
SS03	0.44	0.14
SS04	-0.00	0.14
SS05	1.69	0.14
SS06	0.36	0.14
SS07	0.08	0.14
SS08	1.36	0.14
SS09	-0.78	0.17
SS10	0.42	0.14
SS11	0.60	0.13
SS12	1.32	0.14
SS13	-0.02	0.14
SS14	0.51	0.14
SS15	-0.73	0.16
SS16	-1.73	0.23
SS17	-1.90	0.24
SS18	-1.37	0.20
SS19	2.69	0.18
SS20	0.14	0.14
SS21	1.10	0.14
SS22	1.25	0.14
SS23	0.78	0.13
SS24	0.85	0.13
SS25	0.47	0.14
SS26	0.51	0.14
SS27	-0.12	0.14
SS28	0.25	0.14
SS29	1.42	0.14
SS30	1.12	0.14
SS31	0.65	0.13
SS32	0.99	0.13
SS33	0.47	0.14
SS34	1.27	0.14
SS35	0.67	0.13
SS36	1.78	0.15
SS37	0.31	0.14
SS38	1.38	0.14
SS39	1.01	0.13
SS40	1.97	0.15

Items calibrated on 255 persons

J.7 Difficulty Estimates from Combined Calibration of G1 + G2 + M1(TN)

ITEM NAME	ITEM DIFFIC.	STANDARD ERROR
G101	-1.38	0.27
G102	-1.01	0.24
G103	-0.37	0.20
G104	-0.80	0.22
G105	-0.45	0.20
G106	-1.18	0.25
G107	-0.41	0.20
G108	-0.33	0.19
G109	0.94	0.16
G110	0.81	0.16
G111	1.60	0.16
G112	1.15	0.16
G113	-0.53	0.20
G114	0.26	0.17
G115	0.43	0.17
G116	0.38	0.17
G117	0.81	0.16
G118	0.84	0.16
G119	-1.25	0.26
G120	-1.06	0.24
G121	-0.15	0.19
G122	-0.26	0.19
G123	-0.05	0.18
G124	0.38	0.17
G125	1.57	0.16
G126	1.33	0.16
G127	-0.53	0.20
G128	0.52	0.17
G129	-0.09	0.18
G130	0.17	0.18
G131	0.89	0.16
G132	1.10	0.16
G133	-0.49	0.20
G134	0.20	0.17
G135	1.12	0.16
G136	0.08	0.18
G137	1.20	0.16
G138	1.79	0.17
G139	0.68	0.16
G140	0.86	0.16
G201	-1.18	0.25
G202	0.49	0.17
G203	0.46	0.17
G204	0.49	0.17
G205	-0.90	0.23
G206	-0.26	0.19
G207	-0.22	0.19
G208	-0.66	0.21
G209	0.68	0.16
G210	-0.45	0.20
G211	0.23	0.17
G212	-0.30	0.19
G213	1.65	0.17
G214	0.60	0.17
G215	-0.41	0.20
G216	1.18	0.16
G217	-0.95	0.23
G218	-0.90	0.23
G219	-0.85	0.22
G220	-1.91	0.33
G221	-0.09	0.18
G222	0.52	0.17
G223	0.32	0.17
G224	-0.66	0.21
G225	1.31	0.16

ITEM NAME	ITEM DIFFIC.	STANDARD ERROR
G226	-0.57	0.21
G227	3.77	0.28
G228	0.92	0.16
G229	0.79	0.16
G230	2.55	0.19
G231	-0.57	0.21
G232	0.14	0.18
G233	0.79	0.16
G234	0.94	0.16
G235	2.30	0.18
TN01	-1.18	0.25
TN02	-0.66	0.21
TN03	-2.46	0.42
TN04	-1.91	0.33
TN05	-0.57	0.21
TN06	-2.15	0.37
TN07	-2.03	0.35
TN08	-2.46	0.42
TN09	-1.25	0.26
TN10	-0.05	0.18
TN11	0.38	0.17
TN12	0.46	0.17
TN13	1.28	0.16
TN14	-1.01	0.24
TN15	-1.81	0.32
TN16	-0.95	0.23
TN17	0.11	0.18
TN18	-0.15	0.19
TN19	-0.22	0.19
TN20	-0.41	0.20
TN21	0.79	0.16
TN22	0.60	0.17
TN23	0.26	0.17
TN24	0.08	0.18
TN25	-1.31	0.26
TN26	-0.53	0.20
TN27	0.26	0.17
TN28	0.35	0.17
TN29	-0.66	0.21
TN30	-0.26	0.19
TN31	-0.62	0.21
TN32	-2.30	0.39
TN33	0.49	0.17
TN34	-1.18	0.25
TN35	0.81	0.16
TN36	-0.22	0.19
TN37	-0.15	0.19
TN38	0.57	0.17
TN39	1.54	0.16
TN40	0.60	0.17

Items calibrated on 175 persons

J.8 Difficulty Estimates for G1 from High- & Low-Scoring Subgroups

500 Highest Scorers

ITEM NAME	ITEM DIFFIC.	STANDARD ERROR
G101	-2.88	0.71
G102	-1.80	0.41
G103	0.00	0.18
G104	-1.64	0.38
G105	-0.40	0.21
G106	-0.23	0.20
G107	-0.27	0.20
G108	-0.27	0.20
G109	1.38	0.11
G110	0.77	0.13
G111	2.53	0.10
G112	1.70	0.10
G113	-2.48	0.58
G114	-0.64	0.24
G115	0.17	0.17
G116	-0.03	0.18
G117	1.05	0.12
G118	0.52	0.14
G119	-0.75	0.25
G120	-0.75	0.25
G121	-0.64	0.24
G122	-1.29	0.32
G123	-0.64	0.24
G124	0.77	0.13
G125	2.18	0.10
G126	1.66	0.10
G127	-1.10	0.29
G128	0.73	0.13
G129	-0.09	0.19
G130	0.39	0.15
G131	0.77	0.13
G132	1.28	0.11
G133	-1.64	0.38
G134	-0.53	0.23
G135	0.94	0.13
G136	-0.69	0.24
G137	0.20	0.16
G138	1.35	0.11
G139	0.35	0.15
G140	0.03	0.18

500 Lowest Scorers

ITEM NAME	ITEM DIFFIC.	STANDARD ERROR
G101	-1.78	0.13
G102	-1.57	0.12
G103	-0.64	0.10
G104	-1.13	0.11
G105	-0.89	0.10
G106	-1.24	0.11
G107	-0.95	0.10
G108	-1.10	0.11
G109	0.41	0.10
G110	0.74	0.10
G111	1.04	0.10
G112	0.93	0.10
G113	-0.86	0.10
G114	-0.16	0.09
G115	-0.23	0.09
G116	-0.36	0.10
G117	0.83	0.10
G118	0.22	0.09
G119	-1.46	0.12
G120	-1.21	0.11
G121	-0.23	0.09
G122	-0.65	0.10
G123	-0.19	0.09
G124	0.17	0.09
G125	1.09	0.11
G126	1.47	0.12
G127	-0.82	0.10
G128	-0.06	0.09
G129	-0.15	0.09
G130	0.14	0.09
G131	0.54	0.10
G132	0.72	0.10
G133	-0.12	0.09
G134	0.33	0.09
G135	1.49	0.12
G136	0.35	0.09
G137	1.47	0.12
G138	2.12	0.15
G139	0.59	0.10
G140	1.14	0.11

J.9 Difficulty Estimates for G2 from High- & Low-Scoring Subgroups

500 Highest Scorers			500 Lowest Scorers		
ITEM NAME	ITEM DIFFIC.	STANDARD ERROR	ITEM NAME	ITEM DIFFIC.	STANDARD ERROR
G201	-1.81	0.32	G201	-1.81	0.13
G202	-0.28	0.16	G202	-0.08	0.09
G203	0.71	0.11	G203	0.36	0.09
G204	0.45	0.12	G204	0.58	0.10
G205	-1.01	0.22	G205	-0.78	0.10
G206	-0.41	0.17	G206	0.02	0.09
G207	-0.01	0.14	G207	-0.05	0.09
G208	-0.14	0.15	G208	-0.02	0.09
G209	0.23	0.13	G209	0.33	0.09
G210	-0.05	0.15	G210	-0.12	0.09
G211	0.33	0.13	G211	0.17	0.09
G212	-0.96	0.21	G212	-0.39	0.09
G213	1.59	0.10	G213	1.33	0.11
G214	0.64	0.12	G214	0.83	0.10
G215	-0.96	0.21	G215	-0.59	0.10
G216	0.76	0.11	G216	0.86	0.10
G217	-1.22	0.24	G217	-1.05	0.10
G218	-2.71	0.50	G218	-1.19	0.11
G219	-1.40	0.26	G219	-1.10	0.11
G220	-2.71	0.50	G220	-1.89	0.13
G221	-1.06	0.22	G221	-0.67	0.10
G222	-0.41	0.17	G222	0.20	0.09
G223	-0.84	0.20	G223	-0.16	0.09
G224	-1.01	0.22	G224	-0.66	0.10
G225	1.02	0.11	G225	0.75	0.10
G226	-0.41	0.17	G226	-0.97	0.10
G227	4.87	0.17	G227	1.89	0.13
G228	0.69	0.11	G228	0.61	0.10
G229	0.53	0.12	G229	0.41	0.09
G230	2.07	0.09	G230	1.79	0.13
G231	-0.88	0.21	G231	-0.87	0.10
G232	0.25	0.13	G232	0.09	0.09
G233	0.94	0.11	G233	0.40	0.09
G234	0.62	0.12	G234	0.35	0.09
G235	2.59	0.10	G235	1.48	0.12

APPENDIX K
PUBLISHED RESULTS

Woods, A. & Baker, R. (1985) Item Response Theory.
Language Testing , 2,2, 117-140.

Item response theory

Anthony Woods *University of Reading* and
Rosemary Baker *University of Edinburgh*

A perennial problem for language testers is the need to construct and select test items with 'good' properties. The difficulty lies in the need to assess the properties of items by trying them out on a sample of subjects whose abilities, in turn, it ought to be possible to measure by observing their response to the items. This paper discusses the more important concepts of item response theory (IRT) – a technique, or set of techniques, developed over the last 25 years, mainly by psychometricians. (An application of IRT was discussed in a recent issue of this journal (Henning, (1984).) Basic concepts are introduced and their implications considered by concentrating on the simplest IRT tool, the Rasch (1960) Model.

A few years ago, there was considerable discussion and controversy in Britain concerning the use of the Rasch Model in educational testing in general and it was vigorously attacked by some statisticians – see for example Tall (1981) and Goldstein (1979). However, the misgivings expressed by these, and other, authors related to the use of the Rasch Model for the establishment of item banks which would be assumed to retain stable properties over a long period of time. The model has other uses. Analysis of items via the Rasch Model can complement classical methods of analysis and provide insights otherwise not easily obtained. We believe that the model can supply useful information to language testers. The fact that it may have been incorrectly applied in the past should not be allowed to prevent its being used intelligently now. Models other than the simple Rasch Model will be dealt with briefly later in this paper.

There are two major aspects of the theory to be examined: the development of a suitable measurement model and the analysis of observed responses assuming the measurement model to be correct.

I Measuring ability and difficulty

Let us assume for the moment that we have several test items designed

for use in a large population of subjects which we would like to calibrate on some scale of 'difficulty'. An intuitively reasonable approach would be to define the difficulty of an item as the proportion of subjects who would give the wrong answer to it. This is equivalent to identifying the difficulty of an item with the probability that a randomly chosen individual will not know the correct answer. (For the time being we will assume that a subject who 'knows' the correct answer will provide the correct answer.) It would then be reasonable to *estimate* the difficulty of an item by the proportion of subjects, in a random sample of subjects, who give an incorrect answer. This is exactly what testers have in mind when they quote 'facility values'.

We will return in a moment to the question of estimating item difficulty. Here we are trying to establish a conceptual basis for the direct measurement of item difficulty and the argument will be clearer if we avoid questions of sampling variability. We will therefore assume for the time being that each item we wish to calibrate can be presented to every member of the subject population and the true value of its difficulty ascertained. Even then we may want to disallow items which are so easy that every subject gets the answer correct since they would be useless for discriminating between different subjects. A similar comment applies to items to which no subject can give the correct answer.

Suppose that we have established a large set of items and item difficulties. What then? Presumably we wish to use the items to construct a test which will enable us to order the subjects according to their ability at some task or skill tested by the items. In particular, suppose we have calibrated a set of items on the difficulty scale and have ordered them according to difficulty. Now, *if all the items are assessing the same skill* and the only difference between them is their difficulty in terms of that skill, then item difficulty should be a *transitive* property, i.e. an individual subject who gives the correct answer to the j-th item should answer correctly all the items $i_1, i_2, \ldots, i_{j-1}$. Any individual who does not know the correct answer to item i_k should also fail on i_{k+1}, i_{k+2}, etc. (In practice it never occurs that items are observed to be wholly consistent in this way, a point discussed shortly.)

We still require a means of translating a subject's responses to the items into a score on some scale of ability. Since the perceived abilities of the subjects are determined by the difficulty of the items they get correct and, at the same time, the difficulties of the items are determined by the ability of the subjects on which they are tested, it would be convenient if subject ability and item difficulty could be measured on the same scale. As will be seen later, the Rasch Model provides one way of doing that.

It is useful at this stage to introduce some of the notation we will require in later sections. There does not seem to be a standard notation in IRT and in what follows we will use the notation of Andersen (1980, Chapter 6). Let α_j be the difficulty of the j-th item and θ_i the ability – however it may be measured – of the i-th individual. If the i-th individual gives the correct answer to the first k items and answers incorrectly thereafter we could then perhaps define his ability by $\alpha_k < \theta_i < \alpha_{k+1}$. This statement can be interpreted in a fairly straightforward way. Suppose $\alpha_k = 0.34$ and $\alpha_{k+1} = 0.35$. Then to say $0.34 < \theta_i < 0.35$ simply means that the i-th individual would not belong to the 34 per cent of the population which gets item k wrong, while (s)he *would* be one of the 35 per cent who give the wrong answer to item $(k + 1)$ (and all the subsequent items). He therefore belongs to the *thirty-fifth percentile class* in the ability range of the population.

This is a perfectly adequate definition of ability. However, in order that it should be an operationally useful definition we would need to know how to calibrate the items. Let us suppose that we have a number of unidimensional items of increasing difficulty and that responses to the items are transitive. Suppose that a sample of eight subjects is presented with the items and responds thus (1 denotes a correct and 0 an incorrect response).

Table 1 Responses of imaginary subjects to imaginary items

	Items						
Subjects	1	2	3	4	5	6	7
1	1	0	0	0	0	0	0
2	1	1	1	0	0	0	0
3	1	1	1	0	0	0	0
4	1	1	1	0	0	0	0
5	1	1	1	1	0	0	0
6	1	1	1	1	1	0	0
7	1	1	1	1	1	0	0
8	1	1	1	1	1	1	1
Total incorrect	0	1	1	4	5	7	7

If the item difficulties, α_j, are estimated by the proportion incorrect we would obtain

$$\hat{\alpha}_1 = 0, \quad \hat{\alpha}_2 = 0.125, \quad \hat{\alpha}_3 = 0.125, \quad \hat{\alpha}_4 = 0.5$$
$$\hat{\alpha}_5 = 0.625, \quad \hat{\alpha}_6 = 0.875, \quad \hat{\alpha}_7 = 0.875$$

(The circumflex denotes an estimate.) Now suppose that α_5, the difficulty of item 5, has been underestimated because the sample contains a higher proportion than the population does of subjects who

can answer that item correctly. Then because of the transitivity assumption there will be a high probability that α_4 has also been underestimated (since all those who have answered item 5 correctly will do the same for item 4) and, to a lesser extent, α_3, etc. A similar argument applies to overestimation of difficulties. In other words, the estimates of difficulty of different items will be positively correlated and will depend on the abilities of the subjects who happen to appear in the sample. (Of course, we would not attempt serious estimation of the difficulties with only eight subjects but the problem of correlated estimates will occur in a sample of any size if the 'proportion incorrect' is used to estimate the difficulty of an item.) Into the bargain, a new sample of eight more able subjects would give the correct answer more frequently to all items: the estimates of item difficulty would be biased by the ability of the subjects in the sample. This effect is clearly demonstrated below with real test data in Figure 3a.

Furthermore, it is most unlikely that a set of observed responses will display the consistency of order of difficulty of items shown above. A more typical set of responses might be as shown in Table 2.

Table 2 A typical response pattern

	Items						
Subjects	1	2	3	4	5	6	7
1	1	0	0	0	0	0	0
2	1	1	1	0	0	0	0
3	1	0	0	1	1	0	0
4	1	1	0	1	0	1	1
5	1	1	0	0	1	1	
	etc.						

The item difficulties do not appear to be transitive. Perhaps they are not, they might not all be measuring the same skill and there would then be no reason why every subject should find the items to be in the same difficulty order. On the other hand, we might imagine that the discrepancies may be due to essentially random effects not directly associated with the difficulty of the item. A particular word might awaken an association with something, unrelated to the test, recently read or experienced by a particular subject which causes him to misunderstand an otherwise easy item, and so on. We might therefore suggest that the measuring device comprised of the items being tested is subject to random measurement errors.

The presence of such random errors as a component of an individual's score is generally recognized in other contexts – for example, it is an essential feature of reliability measures. The response of a single subject to the j-th item might then be modelled as

$$Y_j = g(\alpha_j) + \epsilon_j \qquad \text{(A)}$$

where $g(\alpha_j)$ is some function of the difficulty of the j-th item (and would, of course, also depend on the ability of the subject), and ϵ_j is the 'random' error made in measuring the subject's knowledge of the correct response. Many language testers may feel uncomfortable with a model for item response where the 'true' score may take only the values 0 or 1 and the random error take only the values 0 or 1 (for true score 0) and 0 or – 1 (for true score 1). This feeling of unease is probably caused by a statistical education which openly or implicitly assumes that measurement errors always follow a normal distribution. The error ϵ_j in model A clearly cannot have a normal distribution since it can take only two values. However, biologists have been constructing models with 'non-normal' errors and fitting data to them for more than 40 years. One common type of model used in biology known as the 'logistic regression model', which can be routinely analysed by widely available standard computer package, is, in fact, identical to the Rasch Model.

II The Rasch Model

The measurement model (A) for subject response is not sufficiently well specified to be of any use in practice. First, it refers to the responses of just a single subject when several subjects, almost certainly of varying ability, will usually be observed simultaneously and, second, the precise form of the function g which relates the expected response, in the absence of measurement error, to the item difficulty is not given. The first omission is easily taken care of by rewriting the model as

$$Y_{ij} = g(\theta_i, \alpha_j) + \epsilon_{ij} \qquad \text{(B)}$$

where Y_{ij} is now the observed response of the i-th subject to the j-th item, g is now a function of both the subject ability and the item difficulty and ϵ_{ij} is, as before, a random measurement error associated with the response indicated by the subscripts.

A suitable form for the function g is more problematical. We would like it to satisfy several criteria.

1) It should be capable of producing the observed (binary, 0 or 1) data in some more or less plausible way.

2) The values of the θ_i and α_j should be estimable by a theoretically sound procedure.

3) Standard errors for the estimates ought to be provided by the statistical analysis, so that it is possible to assess how close the estimates might be to the true values.
4) There should be some measure of goodness-of-fit of the data to the model.

How to satisfy even the first of these criteria is not immediately obvious. We need a function of θ_i and α_j which will give the answer 0 or 1 for every possible combination of subject ability and item difficulty. While it is possible to construct rather artificial functions which will achieve that, they are rather unappealing and cause problems with all the other criteria. In order to find a satisfactory solution, it will help to formulate the problem in a different way.

Let us consider afresh what a response of either value actually means. What kind of information does it give us? We have admitted the possibility of error in using the response as a measure of a subject's 'ability' (in some rather abstract sense) to give the true answer to an item. If the observed response is 0, we will not conclude necessarily that the subject lacks the ability to answer the question correctly but only that for some reason, when faced by this item on this occasion, (s)he gave the wrong answer. Of course, we will want to assume that *the higher a subject's ability, the more likely (s)he is to give the correct answer* to any item. In this way we can relax the initial requirement that a subject who gives the correct answer to a 'hard' item must also give the correct answer to all easier items. It would follow that *the lower the item difficulty, the more likely it is that a given subject will provide a correct answer.*

The two underlined passages suggest that it may be helpful to formulate a model which describes a relationship, not between the observed binary data and the θs and αs, but which rather relates the values of θ and α to the *probabilities* that the difference data values could be produced, *viz.*

$$p_{ij} = g(\theta_i, \alpha_j) \qquad \text{(C)}$$

where p_{ij} is the probability that subject i gives the correct response to item j, and $g(\theta_i, \alpha_j)$ is a function, i.e. a mathematical rule, which shows how the value of p depends on subject ability and item difficulty. For the moment we will leave this rule unspecified. In a sense this model is too detailed. Our primary interest does not lie in the event that occurs when a specific subject meets a specific item. We wish rather to be able to say something about each item as a whole or each subject as a whole.

Let us instead consider a model of the form

$$p_j(\theta) = g(\theta, \alpha_j) \qquad \text{(D)}$$

or

$$p_j(\theta) = g_j(\theta)$$

which says that the probability of obtaining a correct response from item j depends on the difficulty of the item (α_j) and the ability θ of *any* subject who might be presented with the item. The function g will take different values depending on the value of θ. The graph of $g(\theta)$ against θ and, by association, g itself is usually called the *item characteristic curve,* ICC, of the j-th item. What kind of function should g be?

The value of a probability must lie in the interval (0, 1) and we require here that a subject of higher ability should have a higher probability of giving the correct answer than one of lower ability, i.e. $p_j(\theta)$ or, equivalently, $g(\theta, \alpha_j)$, should be an *increasing function of* θ. There are many functions which have these two properties. The one chosen by Rasch was the *logistic function,*

$$p_j(\theta) = \frac{e^{\theta - \alpha_j}}{1 + e^{\theta - \alpha_j}}. \qquad \text{(E)}$$

(e is a *transcendental number,* like π in the formula $2\pi r$ for the circumference of a circle, and has the value $e = 2.718281\ldots$. The value of e^x, where x is *any* number, positive or negative, can be obtained on a suitable scientific calculator or computer. The relevant function key will be marked either as e^x or exp.). A typical ICC is shown in Figure 1. It can be seen that $p_j(\theta)$ always takes a value in the range (0, 1) and increases with θ.

It is implied by the model E that the value of the function $p_j(\theta)$ depends on the difference $\theta - \alpha_j$, and therefore subject ability and item difficulty have to be measured on the same scale. The ICC has the shape of an elongated letter S – with the 'centre' of the S occurring at the value $p_j(\theta) = 0.5$. Furthermore, if $\theta = \alpha_j$, then

$$p_j(\theta) = p_j(\alpha_j) = \frac{e^{\alpha_j - \alpha_j}}{1 + e^{\alpha_j - \alpha_j}} = \frac{e^0}{1 + e^0} = 0.5.$$

Putting this into words: the Rasch Model (E) implies that a subject whose ability is exactly equal to the difficulty of a given item will have the same chance of answering the item correctly or incorrectly. Equivalently, the difficulty of an item can be defined as the ability of a subject who will have equal probabilities of passing or failing the item. For this reason the item whose ICC is plotted in Figure 1 has difficulty $\alpha = 4.2$.

Although the ICC, $p_j(\theta)$, takes values between 0 and 1 it never actually achieves those values. The model assumes that however able subjects may be, they are never *certain* to give the correct

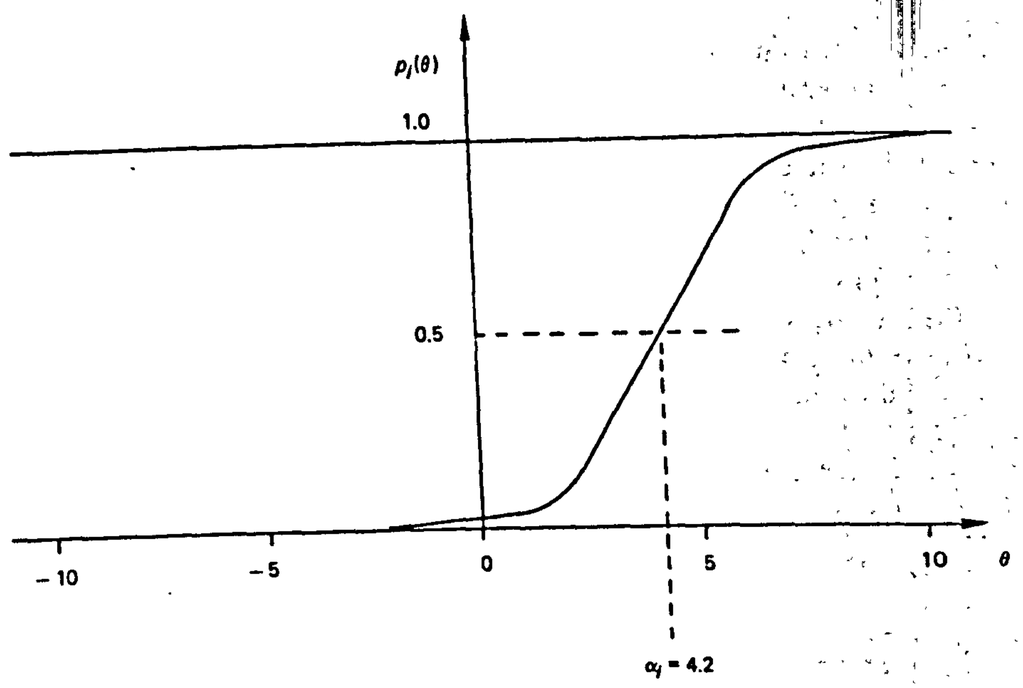


Figure 1 A typical item response curve

answer even to an easy item; similar subjects of low ability will still have a small chance, albeit a vanishingly small chance, of passing a hard item. Subject abilities can take any value on the scale $-\infty$ to $+\infty$, negative scores for less able and positive for more able. Since items can be chosen to match the ability of any subject, item difficulties will vary on the same scale, negative values of item difficulty corresponding to easier items. The ICCs for five items of different difficulties are plotted in Figure 2.

The Rasch Model can be stated in a different form — a form which is in many respects more attractive. From (E) above we have the probability of getting the correct answer as

$$p_j(\theta) = \frac{e^{\theta - \alpha_j}}{1 + e^{\theta - \alpha_j}}$$

and therefore the probability, $1 - p_j(\theta)$, of an incorrect answer is

$$1 - p_j(\theta) = \frac{1}{1 + e^{\theta - \alpha_j}}.$$

Now, the *odds* that a correct answer will be observed is defined to be

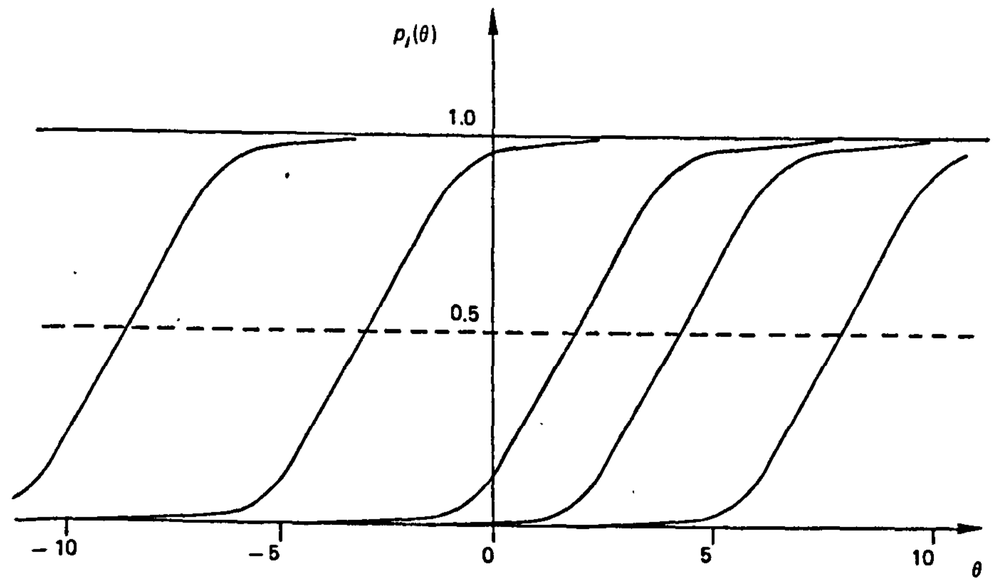


Figure 2 Item characteristic curves

$$\frac{\text{Probability of correct answer}}{\text{Probability of incorrect answer}} = \frac{p_j(\theta)}{1 - p_j(\theta)} = \frac{e^{\theta - \alpha_j}}{1 + e^{\theta - \alpha_j}} \div \frac{1}{1 + e^{\theta - \alpha_j}}$$

i.e.
$$\frac{p_j(\theta)}{1 - p_j(\theta)} = e^{\theta - \alpha_j}.$$

(Apologies to any reader who finds the algebra — especially the step about to come — difficult! However, since the Rasch Model is sometimes stated in form (E) and sometimes in form (F), those who can follow the algebra may find it helpful.)

Now, if we take the *natural logarithm* (i.e. logarithm to the base e rather than base 10) of both sides of this equation, we get

$$\log\left(\frac{p_j(\theta)}{1 - p_j(\theta)}\right) = \theta - \alpha_j. \qquad \text{(F)}$$

The quantity on the left is the logarithm of the odds of getting the correct answer and is called the *log-odds* corresponding to the probability of getting item j correct, measured in units called *logits*. The right-hand side is simply the difference between the subject ability and the item difficulty so that both of these will also be measured in logits. Of course, neither can be directly measured — not even if we had available the whole population of subjects — since the response of a subject to an item is no longer deterministic in the sense that, if it were possible to present the same item repeatedly to a subject and elicit an independent response each time,

the answer he would give would sometimes be correct and sometimes not.

However, in the form (F), the Rasch Model falls into a general, well-understood class of models of the *logistic regression* type. The model (F) can be analysed, and the analysis interpreted, in a way similar to a two-way subjects by items ANOVA in which item difficulties and subject abilities can be estimated simultaneously and independently of one another. First the model has to be amended slightly to allow for the presence of several subjects of, presumably, different ability to

$$\log\left(\frac{p_j(\theta_i)}{1-p_j(\theta_i)}\right) = \theta_i - \alpha_j. \qquad \text{(G)}$$

Various authors have described special computer programs for fitting the model (G) to item response data (e.g. Andersen and Madsen, 1977) but there is no need for such a special program. The model can be fitted using standard computer packages such as GLIM or SAS; the program for carrying out an item response data analysis on GLIM (with a worked example) can be obtained on application to the Journal.

The first purpose of the analysis will presumably be to obtain estimates of the item difficulties. Columns 2 and 3 of Table 3 contain data presented by Andersen (1980) on the results obtained from an analysis of the responses of 1000 candidates for 20 of the easier items from the SAT-test. The second column of the table gives the number of candidates (of 1000) who gave the correct answer to the corresponding item. The third column gives the estimated difficulties.

Notice two details. The first is that the higher the number of candidates giving the correct answer to an item, the smaller is its estimated difficulty, so that *the rank order* of items by estimated difficulty is the same whether the estimation is carried out by assuming the Rasch Model or by calculating the simple proportion of correct responses to the item. Second, some of the item difficulties are positive (i.e. more difficult), others are negative, *and the average estimated difficulty is zero.* (The average value of the estimates in Table 3 is apparently − 0.001 (0.02 ÷ 20) but this is due to rounding the values to two decimal places). This is no accident. The experimenter who analysed the data has deliberately constrained the estimates to have an average value of zero! Before explaining why, let us look at the implications of the constraint. Suppose item 1 was dropped from the set of items and the data

Table 3 Estimates of item difficulty based on Anderson (1980)

Item number	Number correct	Item difficulty estimates (a) First 20 items	(b) New items with link	(c) Complete set	(d) Complete set centred on zero
1	140	+ 2.01		+ 2.01	+ 1.22
2	223	+ 1.37		+ 1.37	+ 0.60
3	239	+ 1.26		+ 1.26	+ 0.49
4	315	+ 0.81		+ 0.81	+ 0.04
5	541	− 0.38		− 0.38	− 1.15
6	537	− 0.36		− 0.36	− 1.13
7	390	+ 0.41		+ 0.41	− 0.36
8	419	+ 0.25		+ 0.25	− 0.52
9	668	− 1.09		− 1.09	− 1.86
10	691	− 1.23		− 1.23	− 2.00
11	77	+ 2.73		+ 2.73	+ 1.96
12	206	+ 1.48		+ 1.48	+ 0.71
13	268	+ 1.08	− 2.30	+ 1.08	+ 0.31
14	425	+ 0.22		+ 0.22	− 0.55
15	696	− 1.26		− 1.26	− 2.03
16	713	− 1.37		− 1.37	− 2.14
17	685	− 1.19		− 1.19	− 1.96
18	720	− 1.42		− 1.42	− 2.19
19	784	− 1.88		− 1.88	− 2.65
20	726	− 1.46		− 1.46	− 2.23
21			− 0.07	+ 3.31	+ 2.54
22			+ 1.28	+ 4.66	+ 3.89
23			+ 0.65	+ 4.03	+ 3.26
24			− 1.40	+ 1.98	+ 1.21
25			+ 1.84	+ 5.22	+ 4.45
Mean difficulty		0.00	0.00	0.77	0.00

reanalysed. The new analysis would estimate the αs for the remaining items and these estimates would necessarily be different from those of Table 3. If they were not they would no longer have a zero average. *The difficulty of an item will depend on the other items whose difficulties are estimated at the same time.*

On the other hand, if this same set of 20 items were tested by presenting them to a new, large sample of candidates, the estimated difficulties would be very similar to those of Table 3, even if the average ability of the new sample of candidates was quite different from that of the original sample. *For large samples of candidates, the estimates of item difficulties will vary little from one sample to another — provided the subjects' responses do follow the Rasch Model.* To satisfy the assumptions of the Rasch Model, items should

be unidimensional and have 'local stochastic independence' which means that the items should not be linked in any way which would cause responses on one item to be related to responses on another. These assumptions are mentioned again in the concluding section.

Why should the estimated difficulties be constrained to have an average (or, equivalently, a sum) of zero? Consider again equation (G). The model defined there says that the log-odds of a correct response depends on the difference between the subject's ability and the item's difficulty. The data provides information about the log-odds of correct response and the analysis then attempts to deduce the values of θ_i and α_j. However, if a subject with ability $\theta_i + c$ (where c is any number) is presented with an item of difficulty $\alpha_j + c$, the log-odds, following the model, will have the value $\theta_i + c - (\alpha_j + c) = \theta_i - \alpha_j$! Changing the ability of the subject and the difficulty of the item by the same amount causes no change in the expected response. Any two sets of estimates in which all the estimates in one set differed from the corresponding estimates in the other by a fixed constant value would fit equally well or equally badly to the data. There would therefore be an infinite number of possible solutions. The conventional way to resolve this kind of indeterminacy when fitting models to data is to impose an arbitrary, convenient constraint. We could, for example, insist that the first item should have difficulty zero. We could insist that the average of the estimated subject abilities is zero (or any other preassigned value, for that matter) or, as is most often done, we set the difficulty estimates to average zero.

The major drawback is that a difficulty value has no *absolute* meaning. It does not convey any information (referring to Table 3) to say that item 13 has difficulty 1.08 except to indicate that it was more difficult than average among its fellows. It *is* meaningful to say that the difference in difficulty between item 17 and item 18 (0.46 logits) is about the same as the difference in difficulty between item 2 and item 3 (0.45 logits) but only experience enables testers to acquire an intuitive feel for what this difference indicates.

III Establishing item banks via Rasch analysis

It is the stability in the difference between item difficulties which is one of the most useful properties of the Rasch Model. It enables the gradual formation of banks of items whose relative difficulties are known, without the need to test all the items in the bank simultaneously. Return to Table 3. Suppose it had been felt that the original 20 items did not contain enough items at the 'harder' end of

the scale. (Perhaps the test was to be used, occasionally at least, with special groups of very able subjects.) Another five items are constructed which are expected to be difficult (items 21–25). These items, *together with some of the items already in the bank,* are then tested on a new group of subjects. There is no hard and fast rule concerning how many of the earlier items should be included. In theory a single item might be enough (and we have used only one such item in our first example) but that would require a very strong belief that all the items had in very large measure the properties stipulated by the Rasch Model. Neither should there be more old items than new items – that would be inefficient. A number between three and ten should be adequate in most situations.

Suppose that the test set comprised items 13, and 21–25. The estimated difficulties are shown in the fourth column of Table 3. Note that the difficulty of item 13 has changed considerably, from 1.08 to – 2.30. You will remember that estimated difficulties are conventionally forced to average zero. Item 13 was of above average difficulty in the first set and so had a positive value. The second set has been constructed deliberately to contain mostly hard items so that, as a member of this set, item 13 is easier than average – indicated by its negative value. However, the *differences* between α_{13} and the other item difficulties are correctly estimated – assuming always the Rasch Model is a good fit! – in both sets. It is therefore possible to combine the two sets of items on a single difficulty scale as follows.

The first step is to give the *link item* the same difficulty in both sets. It will not matter which of the two is chosen though it will be more convenient to leave its value unchanged in the set containing more items, so we will do that. We therefore alter the value for item 13 in the second set to $\hat{\alpha}_{13} = 1.08$. To do this we have added 3.38 logits to the value -2.30. So that the differences in the difficulty estimates remain constant we need to shift the values of $\hat{\alpha}_{21}$, $\hat{\alpha}_{22}$, $\hat{\alpha}_{23}$, $\hat{\alpha}_{24}$ and $\hat{\alpha}_{25}$ upwards by the same amount. All 25 items are now calibrated on the same scale (Table 3, fifth column). However, the mean value of the difficulties is no longer zero; in fact it is 0.77. Since item difficulties produced by a single analysis will always be centred on zero it is often convenient to adopt that as the standard for any set of items. (But see the section below on interpretation of subject ability estimates.) This is easily achieved here by subtracting 0.77 from each of the 25 item difficulties to obtain the final column of Table 3. Thus we have combined the results for different items tested on different samples of subjects. A reader new to this topic may be worried by the use of two different sets of subjects to calibrate the sets of items. One of the most

Table 4 Expanding an item bank using three link items

Item	Difficulty estimates			
	First set	Second set	Combined	Adjusted (c)
1	−2.44		−2.44	-3.165
2	−2.01	(a) −3.50	(b) −2.02	-2.725
3	−1.89		−1.89	-2.615
4	−1.02		−1.02	-1.745
5	−0.60		−0.60	-1.325
6	−0.21	(a) −1.84	(b) −0.29	-1.015
7	+ 0.88		+ 0.88	+ 0.155
8	+ 1.14		+ 1.14	+ 0.415
9	+ 1.64	(a) + 0.35	(b) + 1.73	+ 1.005
10	+ 1.98		+ 1.98	+ 1.255
11	+ 2.53		+ 2.53	+ 1.805
12		−0.53	+ 0.94	+ 0.215
13		+ 1.64	+ 3.11	+ 2.385
14		+ 2.31	+ 3.78	+ 3.055
15		+ 1.57	+ 3.04	+ 2.315
Mean	0.00	0.00	0.725	0.00

(a) The difficulties of the three link items have reduced by 1.47 *on average*. This amount then has to be added to *all* the difficulties in the second set. (b) The link items will each have two (very similar, but different) values: that observed in the first set and that obtained by adding 1.47 to the estimate in the second set. The two values have been averaged. (c) These are the values of the previous column reduced by 0.725 to centre them on zero.

appealing properties of the Rasch Model is that it provides 'sample free' estimates of item difficulties as we demonstrate below.

If the Rasch Model was a perfect fit and the true item difficulties rather than just estimates of their values were known, the procedure just explained would be perfectly satisfactory. In practice it is not safe to use just one link item. The hypothetical example of Table 4 shows why.

There are three link items. The difference in estimated difficulty for the three items is not the same. This is to be expected. The Rasch Model is probably not a perfect fit and even if it were, random variation in responses would prevent complete consistency. It is still simple enough to combine the sets of items using the average difference in the estimated difficulties of the link items. The details are given in the table.

By using link items in this way it is always possible to augment an item bank by incorporating new items whose difficulties can be compared directly with those already in the bank. It will then be possible to develop a set of items which covers the whole difficulty/ability range, to look for new items to fill in any gaps in the range and to prepare large numbers of items at any point in the range which might be of special significance.

Standard logistic regression or ANOVA – and the best of the special programs written to analyse the Rasch Model – also provide standard errors for each of the estimated difficulties. The standard error indicates how precisely the difficulty of the item has been estimated. The standard errors will tend to be large if a) the sample of subjects on which the item was tested is rather small, or b) if the item is much easier or much more difficult than most of those in the group with which it was tested or c) if the test subjects are especially inconsistent. An item which is otherwise attractive but whose difficulty has been rather imprecisely estimated can be tested with a further set of subjects and the estimates combined to give one with a smaller standard error. (This may be one reason for selecting an item as a link item.) It might require the help of a statistician to calculate the standard error of the combined estimate. Of course, every time an item is actually used in a test further information is obtained about its difficulty and, if a bank of items is being maintained, the difficulties of the items could be more and more precisely estimated and any changes in the difficulties (perhaps due to revised teaching methods) can be monitored even as they are being used.

The use of the Rasch Model to construct item banks has been the most controversial issue in Rasch analysis; readers are advised to consult the papers by Goldstein and Tall mentioned in the introduction as well as Goldstein and Blinkhorn (1977) for the anti-Rasch Model argument, and Wright (1977) and Choppin (1981) for the arguments of its supporters. It seems to us that, provided the use of the Rasch Model is tempered by common sense and experience – as *any* item analysis ought to be – and provided it is not intended to create a 'once-and-for-all' bank of items whose properties are expected to remain constant over time, the Rasch Model can be a useful tool for identifying a set of items which cover the part of the ability range in which the tester is interested. On the other hand, no item should be accepted *just* because it seems to fit the model, nor is an item to be condemned *only* because it fits poorly.

IV Interpreting the subject ability estimates

The analysis of item response data under the Rasch Model also provides an estimate of the ability of each subject tested, together with the standard error of the estimate. *Once a bank of items has been established,* the apparent ability of a subject will not depend on which items are included in his test, though the standard error will depend both on the number of items and on their difficulties.

This is an attractive property of the Rasch Model. Note, however, that the meaning of ability = 0 (which is the *same* as difficulty =0!) will change if new items are added to the bank using the standard procedure described above since, as we have seen in the examples of Table 3 and Table 4, the difficulties of all the original items will then be shifted up or down by the same amount if the mean difficulty is to remain zero after the addition of the new items. *An ability of a given value only has a meaning relative to the difficulties of the items in the item bank.* Two subjects, tested on different sets of items from the same bank, who have the same estimated ability, can be considered to have 'equal amounts' of the ability or skill tested by the items. If new items are added to the bank and a third subject is afterwards found to have estimated ability equal to the first two, we must look carefully at how the new items were incorporated in the bank before we will know how to interpret his ability score. In the examples in Table 3 and Table 4, the last step in the calculations adjusted the item difficulties to restore the zero average. While this is the usual convention when forming a basic data bank, it causes a problem when items are being added to an already established bank of items. In order to ensure that the augmented set of items has zero average difficulty, all the original items will have their difficulties changed to new values. If the third subject has been tested on a subset of these items after they have all been increased (say) in value, his ability will be inflated relative to the first two subjects. To give a correct comparison, the abilities of the first two should be increased by the same amount as the difficulties of the original items were.

However, this would create an impossible situation. The measured abilities of subjects would not have a stable meaning. Interpretation of ability values would be impossible without referring them to the content of the bank of items at the time they were measured (not the set of items actually used in the test). The problem can be circumvented by ignoring the last step in Table 3 (or Table 4), i.e. once the basic item bank is established, the difficulties of these original reference items should not be changed when further items are added. Of course, the mean difficulty of the augmented set will no longer be zero, but that has little importance compared with the advantage that difficulty values will not change their meaning, and can eventually be understood intuitively as their continued use makes them familiar. The same will then be true of ability values.

It is sometimes suggested (e.g. Wright and Stone, 1979) that the problem lies in measuring ability on the logit scale because it is difficult to understand what is meant by 'student A has an ability of

1.3 logits'. This is no more true of logits than of any other ability scale. To say that a student has scored 62 per cent of the available marks in a standard test does not convey any information until one has a feeling, an acquired intuition, for the kind of student who is likely to gain that sort of mark in that particular test. Similarly, language testers could soon become clear about the meaning of an ability of 1.3 logits based on a data bank containing a standard set of reference items with known difficulty values.

V The independence of ability and difficulty estimates

Under the assumptions of the Rasch Model, and provided every subject gives a response to every item, the estimates of the item difficulties will not be affected by the abilities of the subjects in the sample used to estimate the difficulties. We will demonstrate this using real data from two different tests designed and piloted by Dr Clive Criper for the Ministry of Education in Malaysia. Both were cloze tests for use in placing learners on a reading scheme as well as for use as a general measure of English proficiency. They were intended to be taken by learners with a very wide range of proficiency, from near-beginner to advanced. Cloze test A consisted of 12 passages containing a total of 141 items, while Cloze B had 13 passages and a total of 147 items. The passages in each case were taken from graded texts and arranged so that the test increased in difficulty from beginning to end. Initially every fifth word was deleted (except for a 10-word introduction to each passage) but some deletions were changed where items were found not to discriminate or where facility values appeared not to be appropriate to the graded design. The 'acceptable word' scoring method was used.

A total of 602 Malaysians spanning a very wide proficiency range were tested on Cloze B. The facility values of the items were calculated twice: the first time from the scores of the highest 200 scorers and the second time from the scores of the lowest 200 scorers. The pairs of facility values are plotted in Figure 3a. (The 15 items with facility values of 0 or 1 for either group are not included.) If facility values were somehow independent of subjects' ability the points should be clustered around the line marked 'identity line' in the figure. Of course, what can be seen is that items found to be easy by the high ability group are found much more difficult by the low ability group so that the points cluster in one corner of the figure.

On the other hand, Figure 3b shows a plot of the difficulty estimates of the same items, again calculated separately for high and

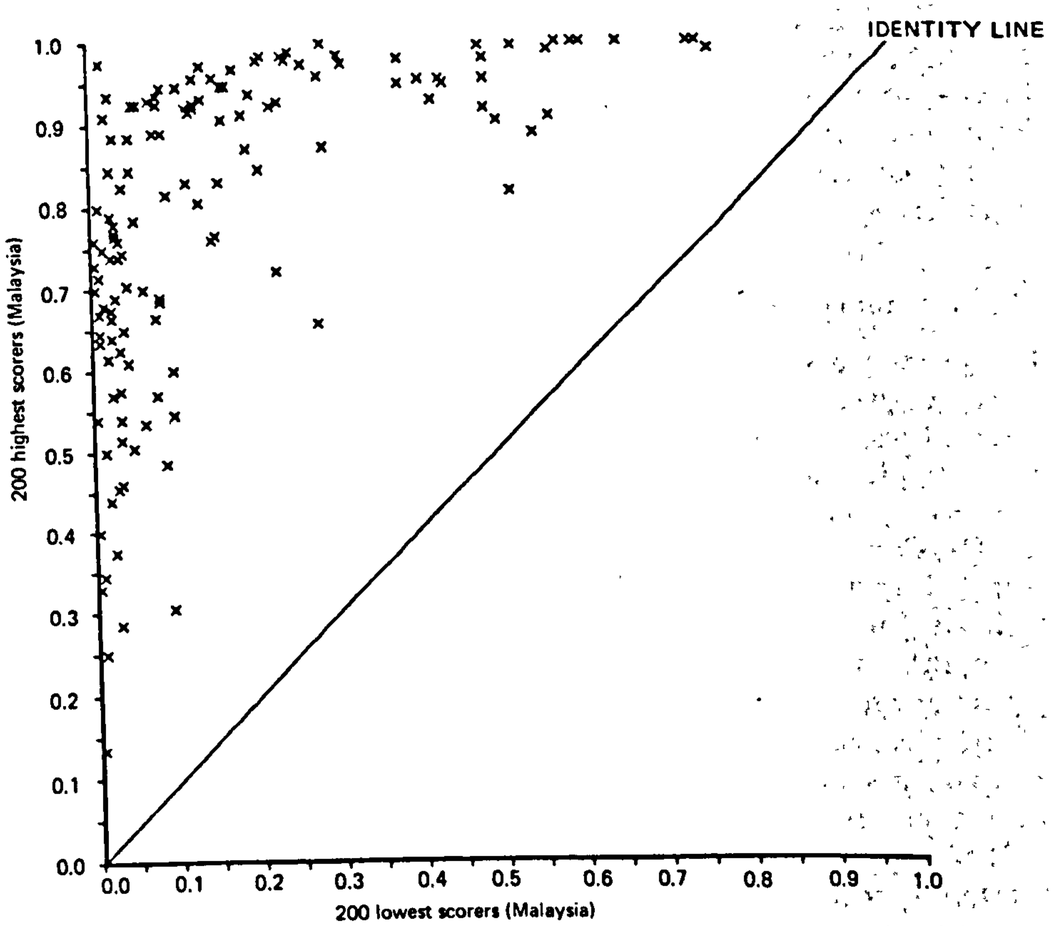


Figure 3a Cloze B facility values: 132 items

low scorers, using BICAL Version 3 written by Mead, Wright and Bell (1979). This time the points cluster around the identity line, demonstrating that the Rasch estimates of item difficulty are not biased in the same way as classical facility estimates by the abilities of the tested subjects. Indeed, it is now perfectly possible that an item can have a larger difficulty estimate from either group. The circled item, for example, is estimated as more difficult from the scores of the higher ability subjects. The items do not fit exactly onto the identity line for two reasons – few items will meet *exactly* the assumptions of the Rasch Model (especially the unidimensionality assumption) and even if they did the model itself permits random errors.

A similar effect is shown in Figure 4 where facilities and difficulties

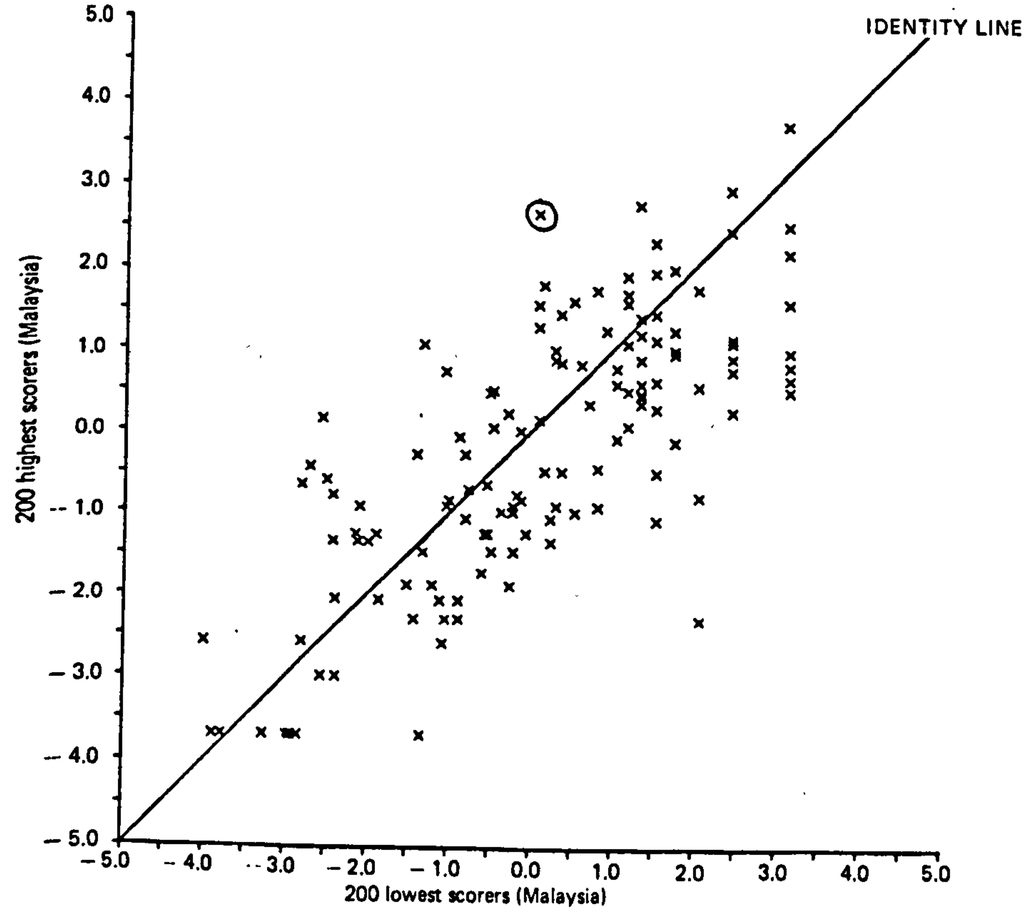


Figure 3b Cloze B, Rasch difficulty estimates: 132 items

were again estimated from two different samples: 611 Malaysians and 243 Tanzanian subjects. Again in Figure 4a it can be seen that one group (the Malaysians) found the items generally easier than the other, while in 4b it can be seen once again that the Rasch Model estimates similar item difficulties from either group.

VI Possible extensions to the Rasch Model

It may well happen that the analysis of a set of item response data indicates that the Rasch Model gives a poor fit. (Wright and Stone (1979) discuss measures of fit.) There are many possible reasons for this, only some of which can be addressed by modifying the basic model. If the poor fit is caused by a high degree of multi-

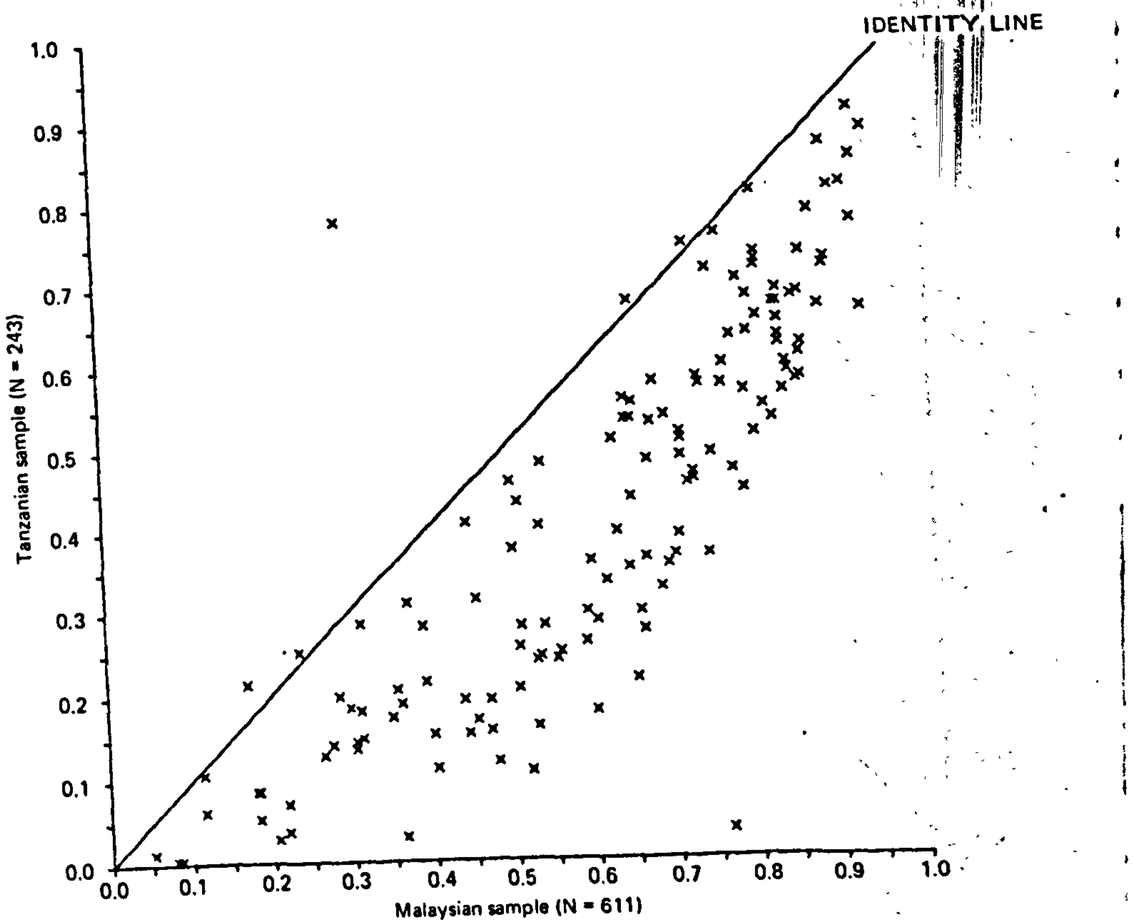


Figure 4a Cloze A facility values (141 items)

dimensionality in the items, if they are measuring ability on different types of skill, then no amount of tinkering with the model will help. (Although, if the subjects' abilities had more or less the same ranking order on *all* the latent skills being assessed, then the basic Rasch Model might still give quite a good fit. This would be true if language competence were a unitary skill.) However, some types of divergence from the basic model can be allowed for, for example:

a ~~*Difficulties*~~ Differences *in item discriminating power*: Graphs of ICCs for items of several difficulties were presented in Figure 2. Although the curves are located at different points on the ability scale they have otherwise the same shape in the sense that they are parallel to one

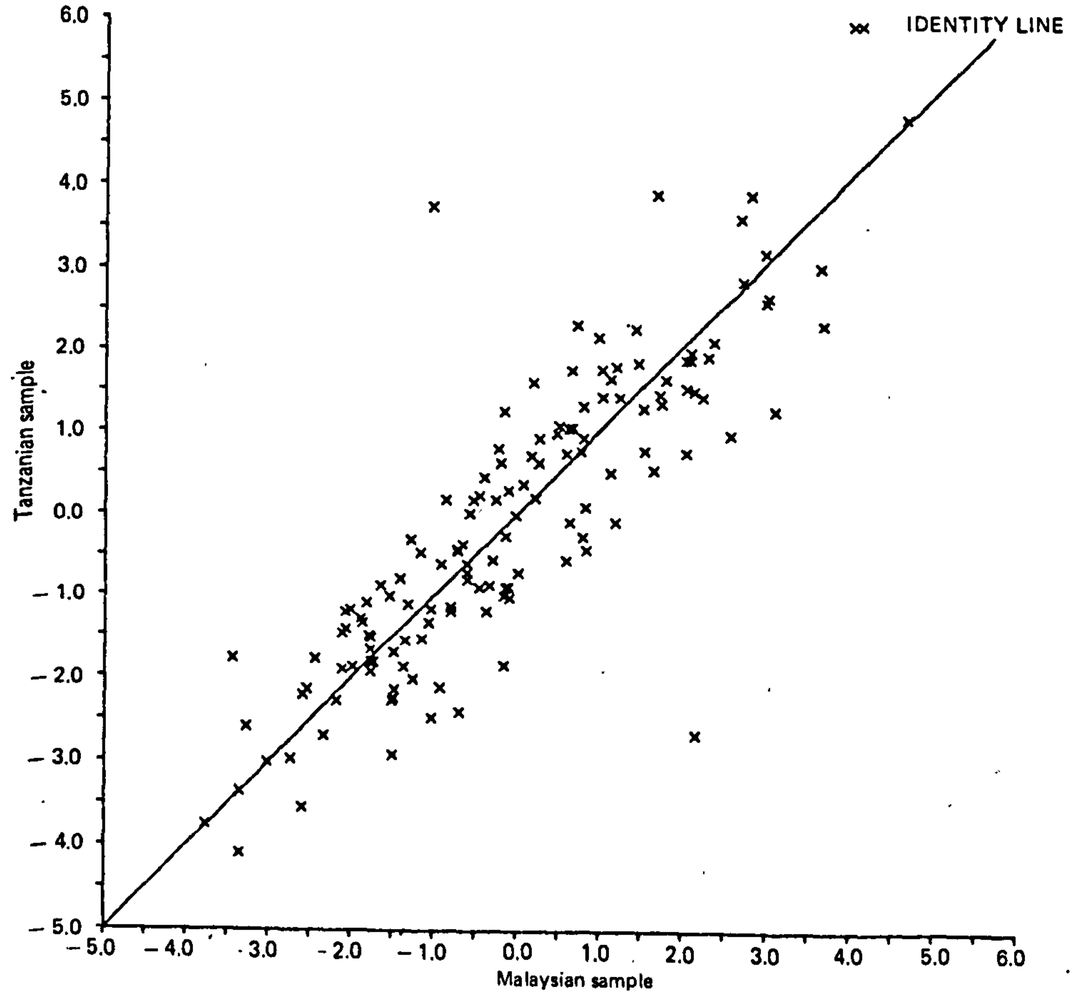


Figure 4b Cloze A, Rasch difficulty estimates (141 items)

another – the basic Rasch Model does not allow the curves to differ except for their location. The two ICCs in Figure 5 are for items with the same difficulty ($\alpha = 1.7$) so that they are both centred on that value, but the ICCs have an important difference. Curve A is much steeper in the centre than curve B. A subject of ability $\theta = 1.7$ will give the correct answer to either item with probability 0.5. However, a subject with $\theta = 0.6$ will almost certainly fail on item A but still has a fair chance of passing item B. A similar remark applies to subjects with abilities greater than 1.7. It may occur in practice that ICCs can differ in this way and, if they do, testers, depending on their aims, may prefer curve A to curve B since subjects with ability only slightly less than 1.7 are almost certain to give an incorrect

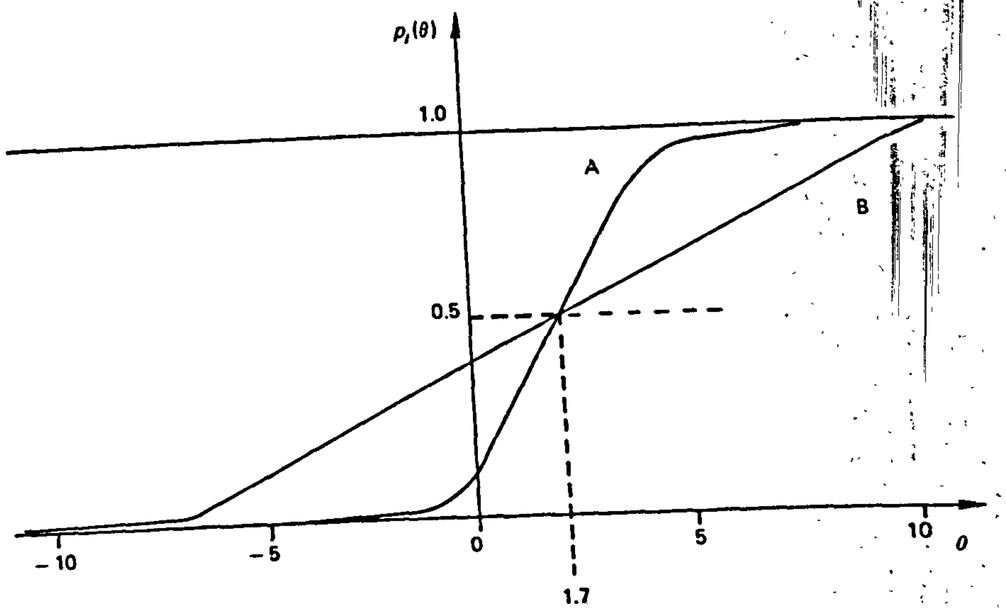


Figure 5 Items with different discrimination power

answer while those with ability slightly greater than 1.7 are almost certain to respond correctly. We could say that *item A has a higher discriminating power than item B*. This idea can be incorporated into the Rasch Model by writing

$$\log\left(\frac{p_{ij}}{1-p_{ij}}\right) = (\theta_i - \alpha_j)\beta_j$$

where β_j is the discriminating power of the j-th item. An example of the use of this model can be found in Lord (1968) and a discussion of the problems of fitting it in Andersen (1973).

b Guessing: Especially with multiple choice items, it is possible for a subject to guess the correct answer to an item. This problem has been discussed by, among others, Lord (1968) and Keats (1974). It is possible to adjust the model of the ICC to incorporate a guessing factor but the model cannot then be put in a form suitable for logistic regression and it will require a special program to fit it to data (see the references for details).

VII Conclusions

The models of item response theory, in particular the Rasch Model,

can be considered as a special case of a type of model which is well-known to statisticians and for whose solution standard software is available. Packages such as GLIM or SAS can be used to obtain estimates of difficulties and abilities on a logit scale together with the standard errors of the estimates.

It is not yet clear how useful IRT might turn out to be in language testing. Items along a single dimension need to be constructed and data banks can be established which contain agreed reference items so that subject abilities can be estimated on a common scale. Testers have to learn how to interpret ability, difficulty and, possibly, discriminating power. In the end, the value of Rasch analysis will depend on how much information testers can extract from it which cannot be obtained at all, or only with difficulty, using classical methods. This is more important than whether it is actually possible to construct items which meet exactly the formal assumptions of the Rasch Model. Users of statistical models in other disciplines have learned to live with 'lack of fit' problems without difficulty. Provided the skill and knowledge of the tester remains paramount there is no danger involved in using any model. We believe that Rasch analysis would provide a useful addition to the techniques already used by testers to assess the properties of test items. We are certainly not advocating that Rasch analysis, alone, can take their place.

VIII References

Andersen, E.B. 1973: A goodness of fit test for the Rasch Model. *Psychometrika* 38, 123–40.

1980: *Discrete statistical models with social science applications.* Amsterdam: North Holland.

Andersen, E.B. and Madsen, M. 1977: Estimating the parameters of the latent population distribution. *Psychometrika* 42, 141–57.

Choppin, B. 1981: Educational measurement and the item bank model. In Lacey, C. and Lawton, D., editors, *Issues in evaluation and accountability.* London: Methuen, 204–21.

Goldstein, H. 1979: Consequences of using the Rasch Model for educational assessment. *British Educational Research Journal* 5, 211–20.

Goldstein, H. and Blinkhorn, S. 1977: Monitoring educational standards – an inappropriate model. *Bulletin of the British Psychological Society* 30, 309–11.

Henning, G. 1984: Advantages of latent trait measurement in language testing. *Language Testing* 1, 123–33.

Keats, J.A. 1974: Applications of projective transformations to test theory. *Psychometrika* 39, 359–60.

Lord, F.M. 1968: An analysis of the verbal scholastic aptitude test using Birnbaum's three parameter logistic model. *Journal of Educational and Psychological Measurement* 28, 989–1020.

Mead, R.J., Wright, B.D. and Bell, S.R. 1979: *BICAL: calibrating items with the Rasch Model.* Research Memorandum 23B, Department of Education, University of Chicago.

Rasch, G. 1960: *Probabilistic models for some intelligence and attainment tests.* Copenhagen: Danmarks Paedagogiske Institut.

Tall, G. 1981: The possible dangers of applying the Rasch Model to school examinations and standardized tests. In Lacey, C. and Lawton, D., editors, *Issues in evaluation and accountability,* London: Methuen, 189–203.

Wright, B.D. 1977: Misunderstanding the Rasch Model. *Journal of Educational Measurement* 14,219–25

Wright, B.D. and Stone, M.H. 1979: *Best test design.* Chicago: Mesa Press.